Cranfield University School of Health

Duane Owen Carey

# The development of novel adjuncts to aid in the diagnosis of Epithelial Misplacement

PhD Thesis

June 2013

Cranfield University

Duane Owen Carey

# The development of novel adjuncts to aid in the diagnosis of Epithelial Misplacement

School of Health

PhD Thesis

June 2013

Supervisors:
Dr Catherine Kendal
Dr Toby Breckon
Prof Neil Shepherd
Prof Nicholas Stone

This thesis is submitted in partial fulfilment of the requirements for the Degree of Philosophy.

Abstract

Epithelial Misplacement (EM) is a benign phenomenon that occurs within polyps most commonly associated with the sigmoid colon. It is brought about because of the colons convulsive nature and this forces a polyps surface epithelium into its submucosa and also causes bleeding.

This is problematic as the Bowel Cancer Screening Programme (BCSP) uses positive Faecal Occult Blood (FOB) test results to identify patients that require pathological review. As EM polyps bleed, they get selected for assessment and this results in them being sectioned and stained. In these cross sections, submucosal glandular tissue will be found that looks like it has formed due to metastatic mechanisms. This can lead to ambiguous diagnoses that will cause some patients to undergo unnecessary surgery.

It is postulated that this can be prevented if the continuity of the EM samples could be measured. This is because only in the EM cases will the submucosal epithelial tissue remain in continuity with the surface. To test this, volumes representative of 9 samples of cancer and 13 cases of EM were segmented and their number of 26 three dimensional (3D) connected components were recorded. These were used with the 99% confidence limits of the two tailed Mann Whitney U Statistic and tested the null hypothesis that the cancer cases were as connected as the EM samples. In this instance, no significant differences were found and so the benefit of measuring the connectivity of these pathologies is questionable.

It was because of this that Immunohistochemical (IHC) alternatives were considered. It was found that Collagen IV antibody staining correctly differentiated nine samples of EM from ten cases of cancer. The Mann Whitney U Statistic found this to be highly significant, $p < 0.001$, and future investigations should concentrate on automating this analysis.

Although, Collagen IV provided a good classification it relied upon the subjective assessment of a pathologist. Therefore, the use of epithelial specific IR spectra was also investigated and this enabled the eleven EM and nine cancer cases that were investigated to be accurately classified 80% of the time upon cross validation. The collection of epithelial specific spectra relied upon a novel digital staining technique that has much application within future research.

This study demonstrates that the intermodal registration of complementary modalities is of benefit to the disease classification problem. This technique has potential to be used in the correct identification of EM but more work is required.

Dedications and Acknowledgements

*I dedicate this thesis to all those who read it.*

I would like to acknowledge Dr Catherine Kendall for her continued support throughout this project and Prof Neil Shepherd and Prof Nick Stone for their expert advice. I would also like to pay tribute to Lynne Bateman, Hazel Clarke and my family as they made completion of this project possible.

# Contents

# List of Figures and Tables

# List of Abbreviations

| | |
|---|---|
| H&E | Haematoxylin and Eosin |
| CK | Cytokeratin |
| IHC | Immunohistochemistry |
| PCA | Principal Component Analysis |
| CPCA | Consensus Principal Component Analysis |
| LDA | Linear Discriminant Analysis |
| MLS | Moving Least Squares |
| scFCM | spatially correct Fuzzy C Means Clustering |
| NCC | Normalised Correlation |
| SIFT | Scale Invariant Feature Transform |
| LOROCV | Leave One Region Out Cross Validation |
| SSD | Sum of Squares Difference |
| EM | Epithelial Misplacement |
| BCSP | Bowel Cancer Screening Programme |
| FTIR | Fourier Transform Infrared Spectroscopy |
| IR | Infrared Spectroscopy |
| PAT | Principal Axis Transform |
| **R** | Rotation |
| *T* | Translation |
| S | Scaling Factor |
| HCA | Hierarchical Cluster Analysis |

# 1.0 Introduction

The aim of the following chapter is to provide a background on the current diagnostic methods used to differentiate cases of Epithelial Misplacement (EM) from cancer. An insight into why current gold standard methods cause significant problems for the Bowel Cancer Screening Programme (BCSP) in the identification of EM along with the epidemiology of colorectal cancer will be provided.

The literature will also be reviewed so that well established and novel classificatory metrics can be identified that have the potential to discriminate EM from cancer.

## 1.1 Epidemiology and detection of colorectal cancer

The diagnosis of colorectal cancer has a detrimental effect on the mental and physical health of patients. However, if it is caught early then the condition can be cured by surgery. This fact is proven in Table 1.1 where over 90% of patients with very early stage colorectal cancers have survived for up to five years after surgical intervention.

**Table 1.1.** A table demonstrating that there is a need to diagnose colorectal cancer early. The Dukes stages are a pathological tool used to grade cancer by its severity [1].

| Dukes' Stage at diagnosis | Percentage of cases/% | Five-year relative survival/% |
|---|---|---|
| **A** | 8.7 | 93.2 |
| **B** | 24.2 | 77.0 |
| **C** | 23.6 | 47.7 |
| **D** | 9.2 | 6.6 |
| **Unknown** | 34.3 | 35.4 |

The progression of colorectal cancers towards more advanced stages is problematic because of the mechanism by which cancer grows. It is thought colorectal cancer develops through the adenocarcinoma pathway, Figure 1.1, and current medical opinion is that invasive cancer only forms towards the end of this sequence [2]. Therefore, it is essential to remove early stage tumours, as their further development will mean that they will gain the ability to metastasise to other regions of the body.

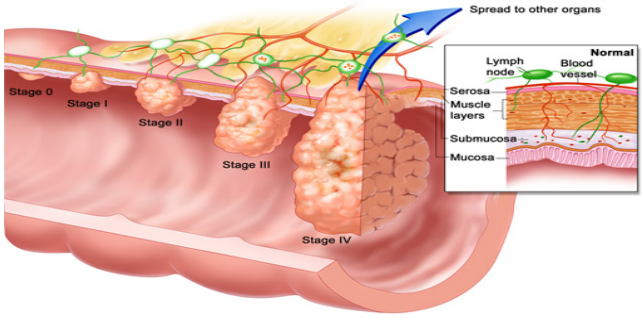



**Figure 1.1.** The adenocarcinoma sequence. The stages refer to the Tumour, Node & Metastases grading (TNM) system[3]. Image sourced from Google images.

To prevent colorectal cancers from reaching their full invasive potential the Bowel Cancer Screening Programme (BCSP) was established in the UK. The BCSP operates via the gold standard of histology and offers high-risk patients a Faecal Occult Blood (FOB) test every two years. This tests for the presence of blood within stools[4], a diagnostic sign of colorectal cancer, but it is not very specific. Therefore, multiple FOB kits are posted to patients and if the BCSP receives two positive results, then they are invited to undergo an endoscopic examination.

A surgeon will remove suspect tissue samples during this time so that they can undergo pathological review. However, before evaluation they are sent to a histology technician who will fix the tissue in formalin and embed it in paraffin. After this they are sectioned and mounted on to glass slides where they are most commonly stained with Haematoxylin and Eosin (H&E)[5].

H&E binds to negatively charged and cytoplasmic constituents respectively and these reagents highlight the morphological structures contained within tissue sections. This allows a pathologist to evaluate the dysplastic properties of samples and if they are regarded as being suspicious, the patient undergoes invasive surgery. If they are classified as normal then the process is repeated in another two years.

This gold standard is usually very accurate and along with improved medication has facilitated a reduction in the mortality rates associated with colorectal cancer. However, the incidence rates for this disease remain high. This is evident from the 412,900 cases of colorectal cancer reported in Europe in 2006[6].

For men, it is the UK's fourth most commonly diagnosed cancer (Figure 1.2) and the disease is more prevalent in men and in older generations [7].

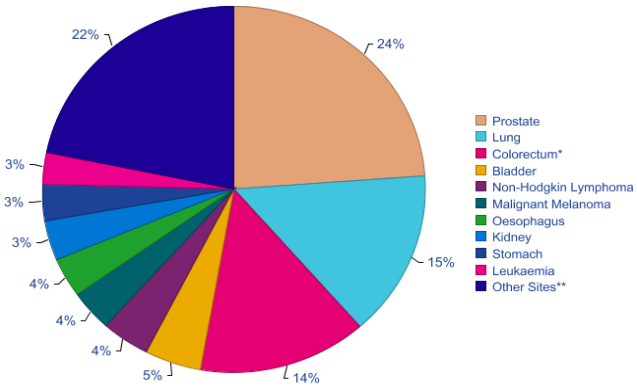



**Figure 1.2.** The types of cancer diagnosed within UK males in 2010. Information sourced from cancer research UK: http://www.cancerresearchuk.org/

Although methods of colorectal cancer detection have improved, the methods of surgical treatment have changed little. The current treatments are extremely invasive and involve removing sections of the colon.

Therefore, any misdiagnosis of a benign condition as cancer will have a huge and negative impact on a patient. Fortunately, H&E staining is very efficient at classifying disease states as this highlights the distinct morphological features that are contained within most pathologies (Figure 1.3). This allows a pathologist's years of training to be used in the diagnosis of colonic disease states.

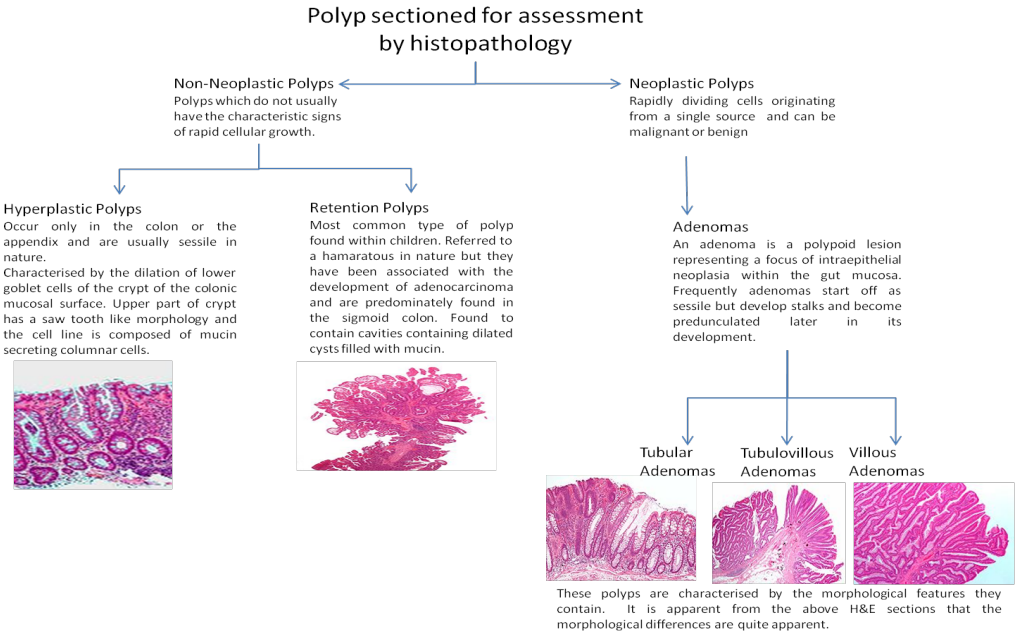



**Figure 1.3.** An illustration of how the different stained features of a colonic polyp can be used in the classification of disease states. The images were retrieved from Google images.

## 1.2 Epithelial Misplacement (EM)

It is evident from Figure 1.3 that there are many distinct colonic pathologies and that visual criteria can be used to differentiate most of them. However, some benign conditions share a homologous morphology with cancer. Therefore, when using the gold standards method of assessment there is a chance that they might get misdiagnosed as cancer. Such a case arises when considering polyps displaying characteristics of Epithelial Misplacement (EM).

EM is a rare phenomenon that occurs most commonly within the sigmoid region[8] of the colon (Figure 1.4). Affected polyps are commonly associated with this region of the colon because of its convulsive nature.

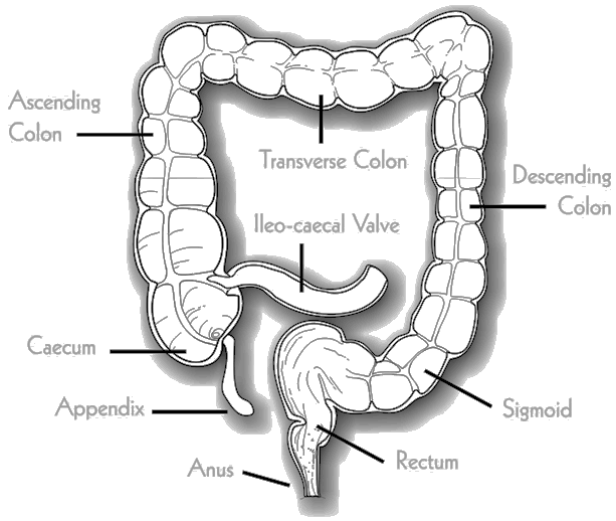



**Figure 1.4.** The location of the sigmoid colon, the place where EM polyps are most commonly found.

The sigmoid colon's movements will cause polyps to become distorted and as a result of this these polyps bleed. The bleeding will cause the FOB test to be positive [4] and will therefore mean that EM polyps get selected into the BCSP. When these samples are processed further, they contain features similar to cancer and this causes patients to have unnecessary surgery.

The similar features found are again a result of the compressive forces that polyps experience within the colon. These forces cause the epithelium from the polyp's surface mucosa to move into its interior and means that submucosal epithelial islands will be found upon sectioning (the orange boxes on Figure 1.5).

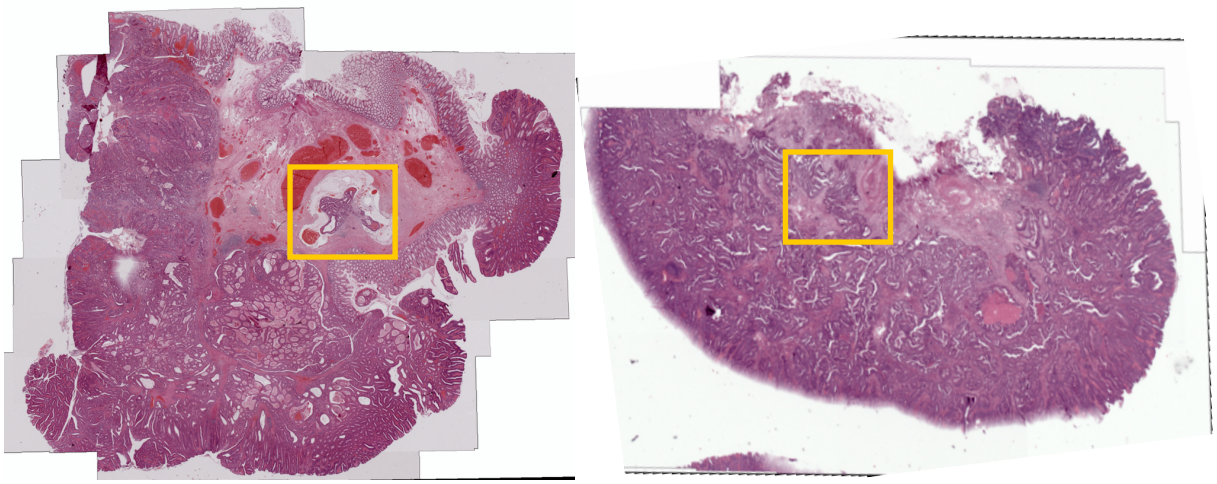


**Figure 1.5.** On the left, a case of Epithelial Misplacement (EM) with the misplaced epithelium surrounded by an orange box. On the right, a case of cancer with the disconnected epithelium bounded by an orange box.

These islands are a characteristic sign of invasive cancer and it is this feature which can lead to ambiguous diagnoses that cause some patients to undergo unnecessary surgery [9].

### 1.3 Current research into Epithelial Misplacement (EM)

To prevent surgery, EM has been traditionally diagnosed using standard histological methods [1], [8] and it has been determined that within EM cases there should be:

1. Gland like structures in the submucosa with the same degree of dysplasia as the epithelium found at the head of an adenoma;
2. Circumscribed submucosal glandular tissue;
3. Hemosiderin deposition in the polyp stalk; and
4. Regular lamina propria around the submucosal epithelial cells.

However, this list is only useful in very obvious cases of EM [9–11] and so new technologies are required. Within the current literature, some studies have adapted standard histological procedures so that discriminatory metrics can be defined. For instance, it has been found that serial sectioning EM polyps allowed this phenomenon's growth characteristics to be analysed. One architectural variable that was found to be useful came from dividing the maximum diameter of the gap in the muscularis propria, which the displaced tissue grew through, by the maximum diameter of the islands themselves [10].

Serial sectioning can also be used with three dimensional (3D) volumes and the literature suggests that this could be used in the identification of EM [9]. It is thought that this technology will be useful because in cases of cancer the epithelial islands have

moved under the influence of their own genetics [8–10]. This leaves behind disconnected islands of epithelium within tissue interiors. This is in contrast to EM where the misplaced epithelial islands are present because of mechanical mechanisms. In EM these islands are thought to remain in continuity from where they were misplaced and if a pathologist could visualise this then EM could be differentiated from cancer [11].

Other current diagnostic adjuncts for EM include antibody staining by immunohistochemistry (IHC). Within the literature, several biomarkers have been identified and the conclusions of the current research are presented within Table 1.2.

**Table 1.2**. The usefulness of identified EM specific antibodies. ~ indicates that some researchers found this antibody useful; + means that all current research agrees on the use of this antibody and – means that the researchers did not find this antibody useful [9], [11–13].

| Antibody | Diagnostic Applicability |
|---|---|
| MIB-1 (Ki67) | ~/+ |
| MMP 1 | + |
| uPAR | + |
| Collagen IV | + |
| E-Cadherin | + |
| p53 | +/~ |
| Stromelysin 3 | - |

The antibodies presented in the above table have been chosen because it is thought that the misplaced epithelium in cases of EM will not have the classical characteristics of invasive tissue. For instance, it is well established that proteins such as Matrix Metalloproteinase 1 (MMP 1), stromelysin 3 and the urokinase-type plasminogen activator (uPAR) system are responsible for the breakdown of the basement membranes which surround the epithelium[11], [14], [15].

The breakdown of this membrane facilitates the metastasis of cancer cells and so these proteins should not be present within cases of EM to a great extent [13]. Other logically chosen markers include Collagen IV. A lack of this protein within the basement membranes of the epithelium indicates fluidity [16] and is associated with metastasis.

Other proteins chosen for analysis by IHC include p53 and E-Cadherin. These proteins are a natural choice as they are involved with the normal regulation of a cell. Loss of function of these proteins can lead to uncontrollable proliferation of cells and the loss

of cell adhesion respectively [17] [18]. However, the studies presented above have usually been carried out independently of one another and on small sample sizes. Therefore, their use still requires further investigation.

Further problems with IHC analysis is that it is very subjective [19–21] and so automated methods are required [13], [19]. One way which the literature suggests that automation can be achieved is through the use of image segmentation and analysis algorithms [19], [20], [22].

For example, an automated system has been proven to produce comparable results to those of two independent pathologists [23]. In this algorithm the size, colour and shape of positive and negative breast cancer cells were determined from training images that had been identified by a pathologist. These were used in the further segmentation of test data and it was noted that the technique could identify positive cells to the same degree of accuracy as a pathologist. The further use of this algorithm will not be dependent upon a pathologist, past the initial training stage, and could be both a valuable diagnostic and educational tool.

# 2.0 New diagnostic technologies available within the literature

The accurate classification of EM will require that efficient diagnostic methods are found. However, the literature available on diagnostic techniques for EM is limited and this is attributable to the fact that it occurs only within a small subset of polyps. Evidence for this can be found in an epidemiological study that estimated that it occurs within between 1% and 3% of polyps [8]. However, preventing the unnecessary surgery associated with this condition has both important economic and physical benefits for patients.

Therefore, the purpose of this chapter was to find methodologies capable of evaluating the potential classificatory metrics that have been previously identified within the literature, such as 3D volume rendering. A description of novel alternative adjunct technologies, such as vibrational spectroscopy, will also be provided. The methodologies contained within the following chapter enabled the hypotheses identified within Chapters Four, Five and Six to be tested.

## 2.1 Image registration using points of correspondence

There is a strong implication in the literature that 3D volume rendering will be able to differentiate cases of EM from cancer [9], [10], [12]. 3D volume rendering involves aligning images of histology sections with respect to one another and this is known as registration.

Image registration can be performed by using either rigid body transformations, which encompass rotations, translations and scaling factors, or by nonlinear means. Linear transformations have a strong presence within the early medical image registration literature [24], [25] but their use with histology images is limited. This is because when a tissue sample is sectioned the appearance of consecutive sections will be naturally different and because a tissue sample can be easily damaged during processing [26], [27].

However, some researchers have successfully applied linear transformations to their volume rendering problems and a discussion of these is warranted here. It should also be noted that the rigid transformation techniques described could easily be extended to aid nonlinear algorithms.

There are many rigid techniques to choose from and a few of these include minimising the Sum of Square Differences (SSD) [28] between images [29]; maximising their mutual information [30] and by using points of correspondence [24], [31], [32].

The first two methodologies cannot be used with images of histology sections because of histological processing and acquisition defects. Typical tissue processing problems include the deformation of tissue samples and staining differences between consecutive sections [24–27], [33]. Image acquisition artefacts include changes in illumination between images of contiguous sections and objects other than the tissue being present within the field of view of the camera [25], [26]. Therefore, points of correspondence algorithms are preferential when considering histology images.

### 2.1.1 Rigid Transformations

The usual application of correspondence points with rigid transformations involves the minimisation of their Euclidean distances. One example of this can be found in a publication by J Streicher et al who used volumetric markers in the linear registration of images of mouse limb sections. However, to do this they deviated away from standard operating procedures and used a resin embedding material [25]. Resin was used so that the impact of damaged tissue sections was limited and it is evident from Figure 2.1. that their registration attempt was successful. Although, their technique is limited to new cases only, as it would be the general case that all retrospective samples will be embedded in paraffin.

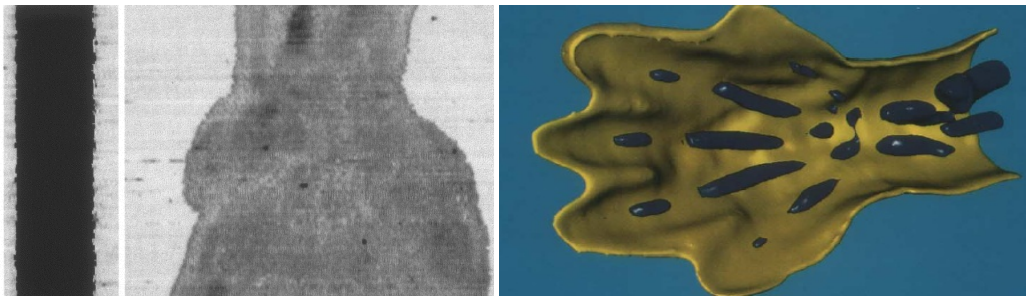


**Figure 2.1.** The volumetric reconstruction carried out by J Streicher et al. The left side represents the robustness of the approach taken and the right side represents the detailed information which can be acquired from 3D models constructed this way [25].

Another extensively used algorithm in the establishment of points of correspondence for linear registrations is the block matching algorithm. It uses similarity metrics to register images [24], [34] and begins by segmenting a floating image into a regular lattice. Blocks from this lattice are then compared to those contained within the local neighbourhood of a static image, the target, via a similarity metric such as normalised correlation (Figure 2.2).

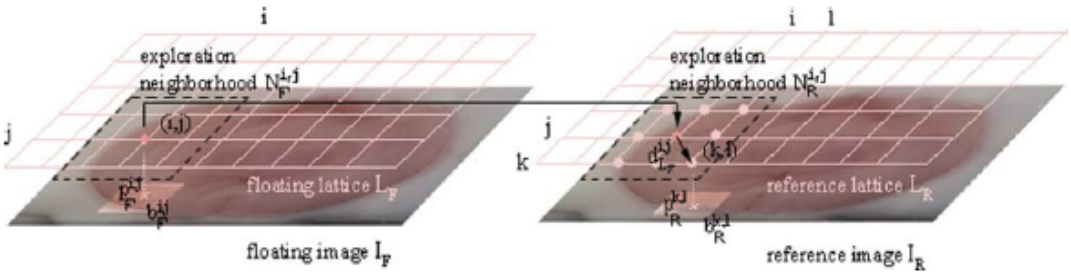



**Figure 2.2.** The block matching algorithm. A block of the floating image, of dimensions NxN, is iteratively moved within the neighbourhood of a reference/static image and a similarity metric is computed. The centres of the blocks with the highest correlation coefficient are then recorded as correspondence points [34].

Maximisation of similarity metrics allows for points of correspondence between the different images to be defined. The choice of the similarity metric is important with normalised correlation being used most commonly within the literature. This is because it is robust to changes in illumination [34]. Normalised correlation is easily computed and is defined by:

$$C(B_R, B_F) = \frac{1}{N^2} \frac{Cov_{(RF)}}{\sigma_R \sigma_F} \quad (1)$$

where:

$$Cov_{(R,F)} = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \left[ I_{R(a+i,b+j)} - \hat{i}_{(a,b)} \right] x \left[ I_{F(u+i,v+j)} - \hat{i}_{(u,v)} \right] \quad (2)$$

In equation 1 & 2 $\hat{i}_{(a,b)}$ refers to the mean of the image intensities in the static image and $\hat{i}_{(u,v)}$ refers to the mean of the image intensities in the floating image. N refers to the number of pixels, I represents the intensity of a pixel at a given location and σ refers to the blocks standard deviation [24].

However, the most current image registration literature has deviated away from block matching and this is because it is computationally cumbersome and rotationally dependent. A multi-iterative approach, as outlined by Ourselin et al, can cope with rotations to an extent but this method still remains unattractive because of the extra computational complexity involved [24].

Therefore, a method that is robust to noise, illumination differences and rotations is required. The Scale Invariant Feature Transform (SIFT) algorithm is invariant to all of these and has been used extensively within the image registration literature [32], [35].

In SIFT, a Difference of Gaussians (DoG) pyramid, which is made up of many levels known as octaves (Figure 2.3), is constructed from pairs of images so that points of

correspondence between them can be identified. This involves searching the levels of the DoG for points of maximum and minimum intensities. These are used to collect pixel orientation and magnitude information from the original images so that they can be included within a feature vector.

Points of correspondence are found from computing the minimum Euclidean distance between these feature vectors and the further minimisation of the spatial distance between these points facilitates registration [35], [32], [36].

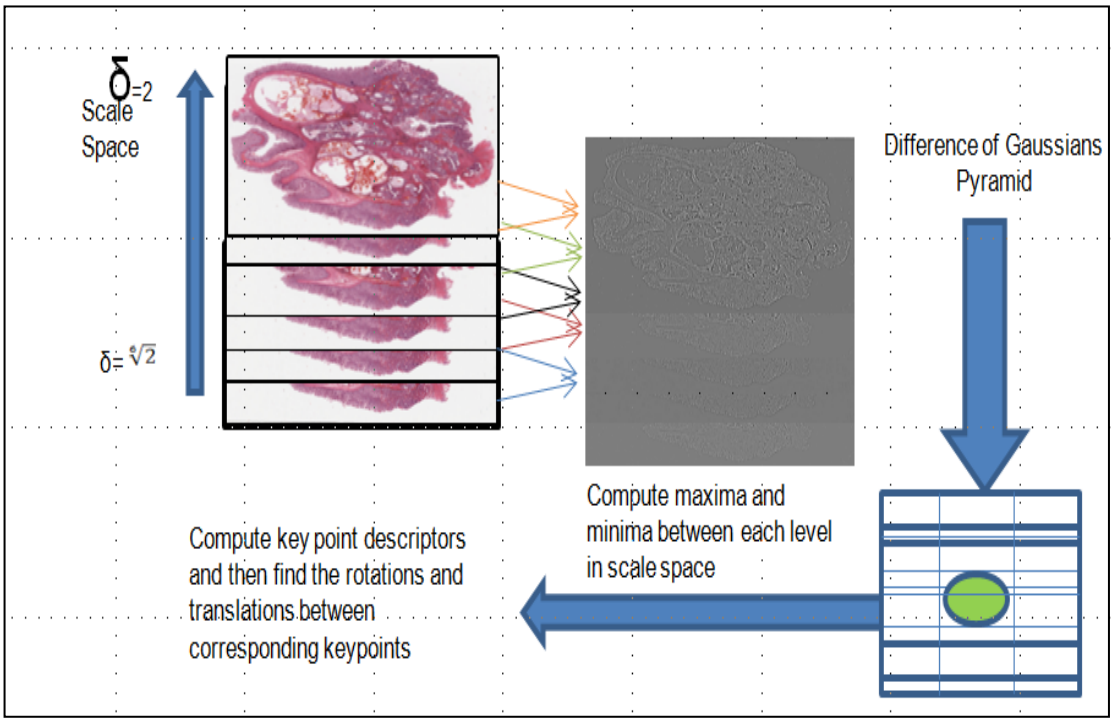



**Figure 2.3**. The SIFT algorithm. It begins with the consecutive convolution of an image with a Gaussian kernel. All convolved images are placed into a pyramid and image pairs are subtracted so key points can be formulated. These key points are used to abstract image information that can then be used in the definition of points of correspondence with another image.

Although the SIFT algorithm is very accurate at matching features between images, it also identifies many outliers. These have to be removed if the points it identifies are to be used in the accurate registration of images. One way in which the literature suggests that this can be done is through the application of the RANdom SAmple and Consensus matching algorithm (RANSAC) [36], [37].

RANSAC refines correspondences by iteratively fitting a model, made up of rotations, translations and scaling factors, to a set of points. This continues until an error threshold has been passed or until the algorithm exceeds its maximum number of iterations. Outliers are classified as those points that do not fit RANSAC's parameters

with a degree of accuracy and this is determined with respect to the rest of the data. RANSAC has been used generally in the literature to estimate affine and linear transformations but can be adapted so it can be used in the non linear registration of images [31], [36].

This is achieved by splitting the population of the putative correspondence points into groups and then applying RANSAC to the subpopulation of correspondences [36]. This non linear adaption is useful because histological tissue sectioning procedures prevent a reliance on linear image registration schemes. This is because of the inevitable deformations that tissue samples undergo.

### 2.1.2 Non linear image registration using points of correspondence

Even though rigid transformations are not optimal, they should not be neglected. They provide a good starting point that non linear image registration algorithms can improve upon. This can be seen many times within the literature, where for instance a Principal Axis Transformation (PAT) was applied before an elastic deformation field was used [26], [38].

Non linear techniques can be generally split into two categories. There are those that can be described as being “as rigid as possible” and those that cannot. In some registration algorithms, such as Thin Plate Splines (TPS), all of the defined correspondence points have an effect on the end registration result[39], [40]. Therefore, the presence of outliers or poor points of correspondence can have a massive effect on the final registration.

In the alternative “as rigid as possible” techniques points of correspondence only have an effect on their local image neighbourhoods. Within the literature there are many too choose from but one state of the art method was developed by the Leeds Institute of Molecular Medicine (LIMM) [41], [42].

This group has access to powerful Aperio T2 (San Diego, USA) scanners that are capable of taking images of entire tissue sections, known as virtual slides, with excellent resolution[43]. In their algorithm, images of tissue sections cut from a block are acquired and then pairs of these images are registered with respect to one another in sequence. To begin with, the algorithm rigidly transforms one of the acquired images to look the one that was taken consecutive to it.

Further to this, points of correspondence are found between the same images. These are a result of comparing tiles from the rigidly transformed image against its consecutive counterpart by correlation [44]. The points identified are used with a B-

Spline grid to enable non linear transformations and the results produced are very accurate. This is evident from recent publications made by this group[41], [42].

Other examples of “as rigid as possible” methods include the Moving Least Squares algorithm (MLS)[31], [45]. This begins by fitting one of the images to be registered with a grid and the vertices of this are moved so that the minimum Euclidean distance between correspondence points can be found. The algorithm is made “rigid” by using weights that are found from calculating the distance of the grids vertices away from the images correspondence points [31]. In this way, vertices that are close to the points of correspondence are affected largely by their displacement and less so by points that are further away.

### 2.1.3 Intermodal registration

The methods of determining points of correspondence within the literature are useful only for the registration of images of a similar type. They are of little use when considering images acquired from different modalities. Intermodal registration is an emerging field and is beneficial to histology because of the extra information multimodal images can provide [46]. Reported benefits of intermodal registration include the improved classification of image objects and the exploration of previously unattainable features [46–48].

The literature is full of examples where images derived from different modalities have been registered together [30] but these acquisition techniques and methodologies are not always beneficial to histology. For instance, in the literature, high resolution Magnetic Resonance (MR) and Computed Tomography (CT) images [49] have been registered together. However, the histological information contained within these is limited. Of more interest to histology, is the chemistry that is associated with pathological features and information of this sort can be derived from Infrared (IR) chemical maps [50].

These are beneficial as traditional histological stains are usually only specific for one type of feature, whereas IR images allow for the visualisation of many. This is because it contains biochemical information and it is thought that the registration of these modalities will facilitate the building of robust classification models [46], [48], [51], [52].

However, it is unfortunate that most of the IR literature is concerned with registering IR thermal images to their Red, Blue and Green (RBG) counterparts [53]. These techniques are not always appropriate for spectroscopic images as they either rely on the minimisation of the distance between manually defined points of correspondence or from maximising the overlap between edge filtered images [53–55].

Although the manual definition of correspondence points will enable registration, the process can be extremely laborious and error prone. Recent algorithms have approximated correspondences by using edge filters [56]. In this method, the closest edge pixels between images are considered as being corresponding. Images are then iteratively warped until the distance between the closest points is minimised. However, the problem with tissue deformation prevents this technique from being reliably applied to histology and IR images.

Other techniques, such as the maximisation of the Mutual Information (MI) coefficient, have also been used to register multimodal images together [30]. However, the different background of the histology and IR images makes the use of MI in the registration of images problematic. The staining process removes paraffin from the standard histology images while the IR sections can be left in their paraffin encasing.

Therefore, the standard RBG images will contain only the slide in its background, whereas the IR image background will contain contributions from paraffin. As a result of this, the background pixel intensities of the IR image will not be as uniform as their standard RBG equivalents. This means that the MI coefficient cannot be easily maximised using the raw images.

## 2.2 Image segmentation

Registration is important as it allows for the overlap of features contained within images of serial sections. However, to extract diagnostic information from these, further processing is required. To derive the features contained within images, segmentation algorithms have to be used and these can be broadly split into two subcategories.

Segmentation can either be performed on each individual pixel within an image or on groups of these. Algorithms that use each pixel in the segmentation of features have a strong presence within the literature but in terms of 3D histology their implementation can sometimes be problematic. This is because there will be vastly different staining colours between contiguous images of tissue sections. However, much success has been had with pixel wise classification methods and so a discussion of these is warranted here.

### 2.2.1 Pixel classification

Within the literature there are a wide variety of pixel segmentation methods to choose from. These range from the simple, such as grey scale thresholding [57], to more complex methods, which involve the multimodal registration of images [58].

In simple thresholding, a value of interest is derived from an image's histogram and all the values below this point are set to zero [57]. Histological components cannot usually be segmented in this way because they are composed of a variety of different intensities. However, they do stain particular colours with certain reagents. Therefore, a colour threshold could be applied and this allows for the retention of specific staining colours [57]. Although, there will inevitably be differences in staining between images and so the use of this methodology is limited.

The differences in colour between images can be attributed to the concentration of the reagents used and will result in over and under staining. However, this is not always problematic because colours can be normalised. One way of achieving this is to use the Hue, Saturation and Value (HSV) colour space [28]. This is an almost linear transformation of the RBG colour space into a cone and is depicted in Figure 2.4. This colour space is invariant to illumination as all the different shades of a colour can be represented by a single value, known as the Hue. The line that runs from the edge of the cone to its centre represents the shade of the Hue and determines how light or dark a colour is.

The HSV space is but only one of the many different colour transformations that are available and others include the L*a*b space. In the L*a*b colour space, L represents the lightness of colours and a and b indicates the transition from green to magenta and blue to yellow respectively. It is a perceptual colour space and is one that approximates the colours that can be seen by human vision. It is similar to the HSV space in that a single dimension, L, within the 3D space represents all shades of a colour and facilitates colour normalisation [28].

Normalisation is important as it allows for similar objects to be represented by similar colours. Theoretically, this will mean that objects within an image can be segmented through the use of a simple threshold. However, manually defining a threshold is both cumbersome and subjective.

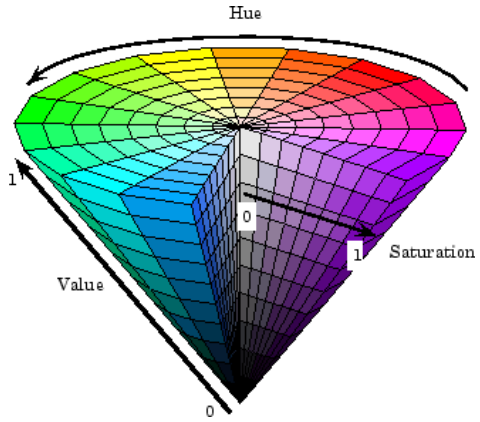



**Figure 2.4**. A diagrammatic representation of the HSV colour space.

It is for this reason that other colour segmentation algorithms are available which provide more automatic solutions. For example, colour deconvolution assesses how much of a particular stain is present within an image [59], [60]. In this methodology, the image's background intensities are used with desired and undesired colours in the retention of objects of interest. This is facilitated by the removal of undesired colour components and allows only the desired stain to be preserved [59]. However, colour deconvolution algorithms still require the manual definition of desired and undesired colour vectors but these can be determined from Principal Component Analysis (PCA) [60].

It should also be noted that, colour deconvolution has application in the balancing of an image's staining colours [33]. This negates some of the issues surrounding differences in staining between consecutive sections and has particular application within 3D volume rendering [33].

### 2.2.2 Texture segmentation

Alternative segmentation methodologies include those that classify image objects by their texture [61]. One well known method is the use of Geodesic active contours with edge filters [28]. In this method, a "snake" is grown on an image and stops when it meets pre-computed edge filtered boundaries, equation 3.

$$E_{total} = \alpha E_{length} + \beta E_{curvature} - \gamma E_{image} \quad (3)$$

In equation 3, $E_{length}$ is the length of the evolving snake, given by a Riemann sum of the line integral calculated over the trajectory of the contour, $E_{curvature}$ is the curvature of the evolving snake (given by a sum of the second derivatives of a snakes position), and $E_{image}$ is the opposing image forces given by the images edge responses. The α,β and γ constants can be changed so the snake can optimally fit the outline of the image object

that it is trying to match. When a snake reaches an images object boundaries equation 3 will be minimised and this stops the growth of the snake.

The problem with the above approach is that the success of the snake in finding the correct contour is entirely dependant on the edge detection method used in the calculation of $E_{image}$. If the rate of change between the boundaries of the images objects is small then the edge filter will not properly represent an images edges and means the snake will not stop growing.

Although, alternative contour methods are available, such as active contours without edges [62]. In this methodology the difference between the mean intensities of the pixels contained inside and outside of an image feature are used to classify objects.

There are many methodologies for evolving contours of this sort, such as those defined by Chan and Vase [62], but the mathematics behind these algorithms is usually very cumbersome. It is because of this that numerical approximations have been made, with the most notable being the growth of contours by the principles of contraction and dilation [62], [63]. This uses a series of rules to determine whether or not the growth of a contour into a particular area is appropriate and can be used with two and three dimensional datasets [63].

Although, active contours are useful in the accurate delineation of image objects they require accurate initialisation before they can be used with histology images. This can be facilitated by implementing the spatially corrected Fuzzy C Means Clustering (FCM) algorithm [64–66]. Hafiane et al used a method of this sort whereby FCM initially represented the overall morphology of images of tissue sections and active contours further refined the results [65].

Other active contour initialisation algorithms used within the literature include Bayesian classifiers [67]. Naik et al used this algorithm to initially classify pixels and the results were further refined using structural parameters and again by the implementation of contours [67].

The use of active contours, or any deformable model, is not the only texture method which can be used to segment objects of interest from within an image [61], [68]. Other methods include the use of neural networks and Gabor filters [61], [69], [70]. Briefly, methods like neural networks operate by an algorithm learning information about an object. They are a convoluted mesh of logical decisions that decide which objects should be segmented [61], [71].

Other methods such as cluster analysis operate by grouping the intensities of pixels by how close they are to the means of randomly initiated clusters. The number of clusters has to be selected before the algorithm operates and so this supervised method of classification is not usually very efficient[61].

Other texture types include Gabor filters. These are convolution kernels which produce responses that allow the intensity gradients of pixels in similar directions to be grouped together [69].

### 2.2.3 Object segmentation by immunohistochemistry (IHC)

Within the literature, conventional immunohistochemical (IHC) antibody staining has also been used to segment histological components. The specificity of these proteins for particular compounds can be exploited so that histology components can be visualised (Figure 2.5).

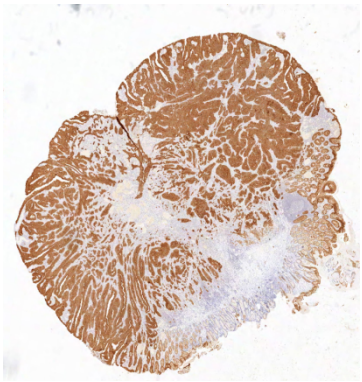


**Figure 2.5.** An example of antibody staining by the MNF116 Pan Cytokeratin (Neomarkers, UK) antibody.

Antibody staining facilitates the identification of proteins because they contain a Y shaped motif, Figure 2.6, that has a conformation that is specific for certain compounds. The matching of the geometry between this site and the molecule of interest facilitates antibody binding [72]. However, the simple binding of tissue proteins to antibodies will not elicit a response as seen in Figure 2.6. For these regions to be highlighted the antibody has to be linked to other molecules[73].

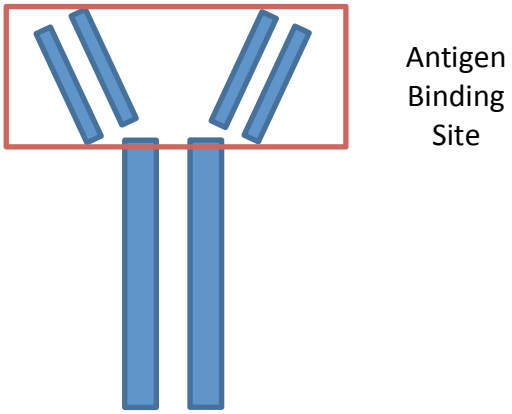



**Figure 2.6**. A graphical representation of the structure of an antibody and its antigen specific binding sites.

Within routine histology the most common method of visualising where antibodies have bound is through the linking of the antibody to a peroxidase enzyme. The peroxidase enzyme catalyses a reaction and causes a tissue section to have a visual appearance that is similar to that of Figure 2.5 [73].

An example in the literature of the use of IHC in the segmentation of histological features was when CD3 was used to highlight the advancing tumour edge of cervical cancer in 3D [58]. This allowed for the diffuseness of the tumour to be evaluated and it was found that it correlated well with pathological assessment.

In other research, collagen has been segmented from within IHC histology images of oral cancers. These fibres were segmented using neural networks and the arrangement of these was used in the differentiation of normal cases from cancer [71]. However, the publicised methodology was very complex and it is evident from their results that features other than the collagenous fibres were being included within multivariate models.

### 2.2.4 Improving segmentation by interpolation

Segmentation is required so that features of interest can be represented but in some instances tissue sections will become damaged or destroyed [27]. This will cause gaps to be found within volumes and so a method of regenerating these is required.

For instance, if the goal is to gauge the continuity of particular histological components, missing or damaged tissue sections will affect the results [27]. Further problems caused by this include the fact that the volume's gaps maybe larger than the voxels that make it up [74]. Therefore, some independent objects maybe generated as artefacts of the tissue sectioning process.

The glandular tissue within the missing sections can be estimated and this could be achieved using a variety of different methods. For example, the location of the epithelium could be approximated by the union of spheres algorithm; linear interpolation methods or by algorithms that use shape based parameters [74], [75]. The union of spheres algorithm transforms segmented image objects into spheres so that features can be matched and interpolated [75]. The use of shape based parameters from segmented objects allows for the improved representation of unknown features but these methods are generally cumbersome and computationally expensive to implement [74].

Other more efficient algorithms, such as the Kalman filter, exist. The Kalman filter has been used previously within the literature both in the segmentation and tracking of tubular objects [76]. This recursive algorithm operates in two stages, referred to as the predictor and the update phase. In these stages, information is taken from an image and allows for the position of an object to be predicted through the use of Newton's equations of motion. The predictive properties of the filter allows it to be used in instances where there is missing data within image stacks and thus facilitates a type of shape based interpolation[76]. However, it only gives information on location and not what the object should look like.

## 2.3 Photonic solutions for the diagnosis of Epithelial Misplacement (EM)

The methods discussed above for the derivation of continuity and IHC classificatory metrics are not the only way in which EM can be diagnosed. For instance, the spectroscopic characteristics of EM polyps, still in the body, could be sampled using techniques such as fluorescence inducing lasers with endoscopy and other lasers in general[77] [78], [79]. However, the depth of the epithelial islands within cases of EM may prevent the clinical use of these techniques.

The depth of the interior epithelial tissue might also prevent other techniques such as Optical Coherence Tomography (OCT)[78], [80] and Optical Projection Tomography (OPT) from being used [81].

OCT, and other three dimensional (3D) modelling techniques, maybe able to identify cases of EM as they allow the connectivity of the epithelium to be graded. However, they are not the only volume rendering techniques available and others include four dimensional (4D) vibrational spectroscopy models [82], [83].

The problem with 4D spectroscopic models is that it takes a long time to acquire images of a suitable quality and a lot of computational memory is required to store the data recorded.

### 2.3.1 Vibrational Spectroscopy

However, vibrational spectroscopy should not be disregarded as it is well cited within the literature as being capable of differentiating disease states [48], [50], [84–90]. This analytical technique provides a label free method of diagnosing disease states and it can be predominately split into two types, Raman and Fourier Transform Infrared (FTIR) spectroscopy. The two methods differ in how they generate spectra with Raman using inelastic scattering [79] and IR using absorption [91].

Vibrational spectroscopy can characterise molecules because their molecular bonds will have unique electronic configurations. These configurations will determine how energy is absorbed/scattered and causes the molecules to have distinct and specific spectroscopic properties[50], Figure 2.7.

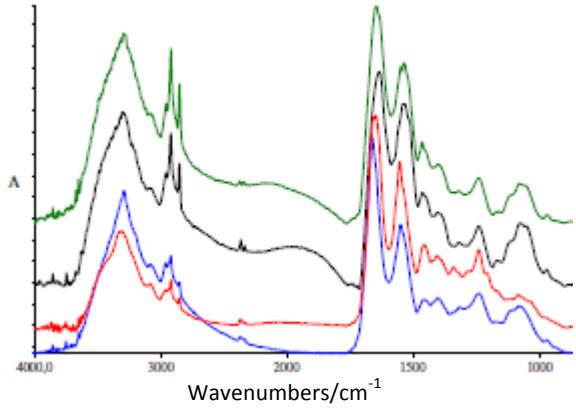



**Figure 2.7.** IR spectra recorded for different tissue pathologies [92].

It is thought that vibrational spectroscopies will have application in the differentiation of EM from cancer because of the different mechanisms by which these conditions arise. In cancer, the metastatic properties of its associated cells will produce different molecular compounds to those of EM. Since they are composed of different molecules they will have different spectroscopic responses and these can be used within multivariate classification models.

### 2.3.2 FTIR spectroscopy

Although Raman imaging has been used previously with colonic tissue [92], the time that is required to record data means that this technique is not always suitable. However, recent technological advances means that this technique has greater applicability within a clinical setting [93].

The time required to acquire Raman images probably explains why more of the literature has used Infrared (IR) technologies in the probing of colonic tissue samples [84], [85], [92], [94], [95]. It is thought that IR will be of real benefit in the reduction of

the workload of pathologists and will also allow the accuracy of pathology to be increased.

Histological processes are regularly scrutinised within the literature because the agreement between different pathologists is highly variable. J Eaden et al utilised kappa statistics to prove that the consensus agreement between a group of pathologists for four different types of colorectal pathologies was inconsistent and unreliable.

In their study, 51 tissue slides representative of varying degrees of Ulcerative Colitis (UC), low grade dysplasia, high grade dysplasia, reactive hyperplasia and inactive colitis, were sent to 13 different pathologists for their assessment. The pathologists chosen had a range of professional experience, with one group being more junior than the other. Their research would suggest that both groups were as equally bad as each other on deciding what a particular pathology was. The junior and professional pathologist scored a kappa value of 0.30 and 0.28 respectively and this indicates a large degree of intergroup variation [96].

In another more recent study, two epithelial misplacement samples, part of a much larger cohort of pathologies, were given to a group of pathologist for their assessment. One sample was given the correct classification by all clinicians and the other sample was completely misdiagnosed. This highlights the need for a more objective technique to be found that can aid in the diagnosis of EM [97].

IR will help because it enables the professional assessment of a few pathologists to be extended too many clinical sites. It also provides a compact and efficient method that can acquire data in clinically relevant time scales [50].

The above advantages are made possible because of the machinery associated with FTIR spectrometers. They advanced greatly over the past few decades and they are now capable of simultaneously sampling all IR frequencies with greater accuracy[91]. These advantages are made possible because FTIR spectrometers contain a Michelson interferometer (Figure 2.8).

This interferometer uses a moving mirror to bring about the interference of two IR radiation beams that have been split and recombined at a beam splitter. This enables the production an interferogram [98] which contains all IR frequencies that are passed through a sample. The detector records the absorbed frequencies and a Fourier Transform enables for the spectroscopic characteristics of a tissue sample to be understood.

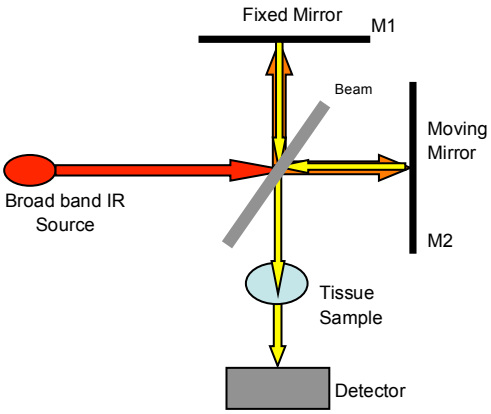



**Figure 2.8.** An illustrative representation of how an interferometer works [99].

When using IR there are many ways of collecting data, such as by point and array mapping, but the most clinically relevant is undoubtedly the former of these. Array chemical imaging involves the use of stage mounts and these enable all of a tissue section to be sampled with IR. False colouring of the raw data results in the production of an image where the chemical differences between pixels have been highlighted and makes the interpretation of IR data easier [98].

#### 2.3.2.1 Methods of IR data collection

The array detectors in IR spectrometers can either be focal or linear in nature but the literature prefers linear systems. They are a good trade off between the high fidelity of point mapping techniques and the faster acquisition times of the focal plane detectors [98].

The options presented above are not the only methodological considerations that need to be taken when acquiring IR data. Other particularities associated with IR include the sampling mode and this can either be done through the use of transmission or transflection spectroscopy.

In transmission, the detector is below the tissue sample and the spectrum recorded is representative of the IR spectra that has passed through the sample and has not been absorbed. In transflection, a double pass is made through the tissue sample and increases the recorded signals strength [99]. However, Mie scattering can sometimes prevent the accurate recording of spectra in this mode. This phenomenon is due to changes in the refractive index between samples and their surroundings. However, techniques, such as the Extended Multiplicative Scatter Correction (EMSC) algorithm, have been designed that can circumvent this problem [100]. EMSC also has applications within other areas and it can be used to improve the quality of the recorded dataset.

### 2.3.3 Improving the data recorded by FTIR spectroscopy

EMSC is preferable to other spectral refinement techniques because other methods such as polynomial fitting [85], [94] will add undesired variation into the recorded dataset. The benefit that the field attaches to the EMSC algorithm is evident from the current literature [50], [88], [100–103]. For instance, recent research has indicated that breast micro metastases have been made more distinct from their surroundings when using this method[101].

Further uses of EMSC include the correction of interfering spectral elements that are related to signal contributions that arise from undesired compounds such as paraffin [102], [104]. In the literature, paraffin has been noted as obscuring important classificatory IR peaks and so a viable method of removing it is required [102], [105].

The EMSC algorithm can be used for this because it assumes that spectra can be decomposed into additive and multiplicative contributions. The additive elements arise from chemical compounds and physical interferences and the multiplicative contributions are introduced by scattering phenomenon [106], [107]. Knowledge of this allows for the definition of a simple formula that can be used to reconstruct the spectrum of any chemical constituent, and is given below:

$$\boldsymbol{s_i} = a_i\hat{\boldsymbol{s}} + \boldsymbol{b_i}I + \boldsymbol{c_i}P + \boldsymbol{e_i} \quad (4)$$

In the above, equation s is the $i^{th}$ spectra under consideration, $\hat{s}$ is a reference spectrum, I is an interference spectrum, P is an nth order polynomial and e is the residual of fitting the $i^{th}$ spectra with coefficients a, b and c. The a, b and c coefficients are found through a least squares fit of the original data with an average tissue spectrum, a interference spectrum and a nth order polynomial respectively. The reference spectrum can be used to correct the recorded spectra by computing the following formulae (in this form the nth order polynomial term is ignored):

$$\boldsymbol{s}_{corr} = \frac{\boldsymbol{s}_i - \boldsymbol{c}_i}{\boldsymbol{b}_i} \quad (5)$$

The nomenclature of equation 5 is the same as equation 4. Equation 4 can be used to remove paraffin because those spectra with a high α coefficient will be more similar in nature to tissue spectra [102], [104]. Those with low α values can be removed and enables the affect paraffin has on multivariate analysis to be reduced.

However, other chemical methods of paraffin removal, such as by xylene and cyclohexane, also exist [105]. In the literature, the chemical removal of paraffin from tissue sections has been reported as being inefficient and causes the chemical

alteration of tissue samples [89]. Therefore, digital means of correcting spectra are preferential.

Paraffin contamination and scattering is not the only problem that can be encountered when recording IR data. When recording spectra from large samples, the size of the data can also be problematic. This problem can be resolved through the use of compression algorithms and one efficient method of doing so is via Self Organising Maps (SOMs). SOMs are a type of neural network that allow all of the spectra within an image to be represented by a reduced number of SOM components. The analysis of the SOM units then allows for more efficient data processing [108].

### 2.3.4 Infrared spectroscopy in the discrimination of disease states

The purpose of refining IR spectroscopy is to ensure that the characteristic differences between pathology groups can be exploited. Much research has been devoted to the use of spectroscopy in the discrimination of disease states and a recent example of this is by A Zwielly et al. Their study used 230 samples made up of normal, mild, serious polyp and cancer polyp pathologies from 78 patients to identify potentially useful biomarkers [94]. They found that the carbohydrate region of the IR spectrum could be used to differentiate colonic pathologies and it was thought that this was related to the increased metabolic activity of cancerous cells. They also found that the phospholipid peak found at 1740$cm^{-1}$ wavenumbers was different between cancer and normal pathologies [94].

Other studies, that have used IR spectroscopy in the differentiation of colonic pathologies, include those by Ramesh et al [85], [95]. Some of the spectroscopic biomarkers they found to be capable of classifying cancer were found at 1025 and 1045$cm^{-1}$ wavenumbers. The $CH_2OH$ and C-O vibrational modes of carbohydrates are the functional groups responsible for these absorptions and this was again attributed to the metabolic profile of cancer.

They also used the signal intensity ratio of the peaks at 1030 and 1080$cm^{-1}$ wavenumbers, due to glucose and phosphate respectively, and the peaks at 1121 and 1020$cm^{-1}$, due to RNA and DNA, for classification. When they used Linear Discriminant Analysis (LDA) to derive functions that would automatically differentiate disease states they wrongly classified 13.8% of normal tissue samples as cancer and 8.4% of cancer cases as normal tissue [85]. It was thought that this was attributable to the small number of samples they included in their study.

The effective differentiation of colonic disease states has also been brought about through the targeting of specific histological features. For example, the spectroscopic signals relating to mucin have been used in the classification of normal and cancerous

samples [88]. This molecule allowed for the accurate differentiation of nine adenocarcinomas and seven non neoplastic tissue samples [88]. However, the differentiation criteria used by Travo et al is unlikely to be repeatable within a clinical setting. This is because K means clustering was used with IR images in the localisation of a tissue samples mucinous components. The amount each mucin cluster covered in each IR chemical map was recorded as a percentage and this was used to classify tissue samples.

Analysis of this sort is problematic because it is restricted by the size of the mucin area measured. For example, if only half a mucin pool was sampled with IR then the discriminatory chemical components might be located in the other half. This would require further biochemical assays to be conducted so that the discriminatory molecules could be better understood.

However, IHC studies have found pathology specific mucin expression profiles. This adds credence to the targeting of mucin with IR. It is thought mucin will be a good target because it is a ubiquitous compound found within the colon. There are nine human derivatives [109–113]. These are known as **MUC1, MUC2, MUC3, MUC4, MUC5AC, MUC5AB, MUC6, MUC7** and **MUC8**.

The mucin proteins are primarily composed of a polypeptide chain, a core carbohydrate linker, a backbone oligosaccharide and a terminal carbohydrate chain [109–111], [114]. Immunohistochemistry studies have indicated that within conventional colonic carcinomas there is an up-regulation of **MUC1** and an associated decrease in expression of **MUC2** (Figure 2.9). It is thought that there is increased **MUC1** expression in colonic carcinomas because **MUC2** is made by mature goblet cells. They are not thought to be present in cancer and so there should be a reduced expression of this protein in these cases [109–111], [114].

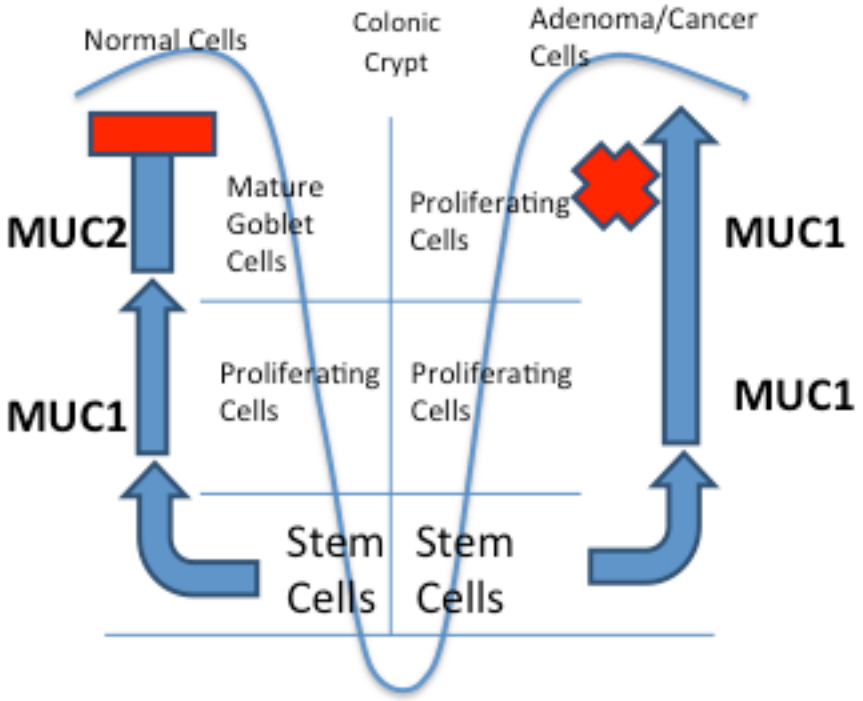



**Figure 2.9.** A theoretical explanation of why there is increased MUC1 within colonic carcinomas.

For other tissue types, such as mucinous carcinomas, the literature has identified a different expression profile, whereby both MUC2 and MUC5AC proteins are up-regulated[115]. These protein types are synthesised in the goblet cells of epithelial tissue and their dominance within particular pathologies has led some researches to believe that this may be indicative of a problem with the TGF-β signalling pathway. This regulates the synthesis of the mucin core proteins in this cell type[110].

Other histological components, such as the epithelium, have been segmented from IR images and used further with multivariate statistics. This involved manually collecting spectra from IR images that had been registered to their white light counterparts [51]. These were then used within a PCA fed LDA model and cross validated using a Leave One Gland Out Cross Validation (LOGOCV) method. The sensitivities and specificities achieved for five different pathology groups, representative of normal, hyperplastic, adenoma, adenocarcinoma and ulcerative colitis, was over 90% [51]. It is therefore evident that collecting histology specific spectra can greatly improve the results of classification.

Other hyperspectral technologies have also been processed in a similar way. Although, in these studies, the morphology of a tissue sample's epithelium was used to classify pathologies instead of the spectral information they contained. However, the spectra were used to correctly identify glandular regions so that a multivariate model could be

parameterised using morphological descriptors. Techniques such as this have allowed normal and cancer cases to be differentiated with classification being in between 75% and 90% [116–118].

## 2.4 Summary of the literature

From the above, it is evident that there is a strong belief that visualising the continuity of glandular tissue within EM cases will enable for its classification [9], [12]. To bring this about, image registration methodologies [29], [31], [34], [36], [40–42], [119], [120] will have to be used with histology images so that 3D volumes can be made. Subsequent to this, image segmentation algorithms will have to be implemented as they facilitate the representation of specific histological components [58–60], [62], [63], [65], [67], [70], [71]. Further, analysis of the continuity of the epithelium will allow for efficient classificatory metrics to be derived that will prevent the unnecessary surgery associated with EM cases.

However, 3D volume rendering is not the only technique that should be investigated. This is because IHC studies have found a host of biomarkers that bring about the discrimination of EM from cancer [9], [11–13]. The automation of this analysis is required [22], [121] but of more importance is the evaluation of the literature's current findings. This will further support the development of larger studies that use similar antibodies.

IHC analysis only allows for the subjective identification of biomarkers and so objective methodologies are required. Vibrational spectroscopies provide a unique solution for this and recent studies have demonstrated the potential of these techniques [50], [51], [79], [84–88], [93–95]. Although, many studies have enjoyed much success with classification by spectroscopy, there are many considerations to take before a technique of this sort is implemented. For instance, the former chapters demonstrate that IR mapping is of more use with colonic tissue than Raman spectroscopy because of the time constraints associated with the former technique [91], [93], [99].

The literature has also shown that IR mapping is preferable to other sampling modalities as this allows for the reliable derivation of histology specific spectra [51], [122]. Similar methodological approaches have been used previously within the literature and they have enabled the accurate classification of similar colonic disease states [51], [122]. Therefore, IR should have application in the differentiation of EM from cancer.

The next chapter will provide methodological details on the novel techniques used to investigate the use of IR, 3D volume rendering and IHC in the differentiation of EM

from cancer. This will be followed by a results section, where the finding of each method is presented, and will end with a section on conclusions and future work.

## 2.5 Project Aims and Objectives

The aim of this project was to evaluate the literatures current stance on the discriminative capabilities of measuring the continuity of epithelial structures and Immunohistochemistry (IHC) staining in the differentiation of EM from Cancer. In this research, 3D volume rendering will be used to assess the potential of epithelial continuity and the work by Yantiss et al will be emulated to review the already published IHC data.

Further to this, a new novel technology, Infra Red (IR), which has not yet been considered within the literature, will also be investigated. In this instance, a variety of different histological targets and sampling methodologies will be considered so that spectral marker points can be found that will help build the basis of larger studies.

# 3.0 Methods

The Ethical approval obtained for this study was from the South West Research Ethics Committee on the 06th of May 2010. The image processing and data analysis for this project was conducted using MATLAB R2007A (Natwick, USA) on a standard desktop personal computer equipped with an I7 Quad core 3.60 Hz processor and 8Gb of RAM.

The histological processing of tissue samples was conducted independently by members of staff working in the pathology department of the Gloucestershire Hospitals NHS Foundation Trust (GHNFT) in Cheltenham. All tissue samples used throughout this project were derived from the GHNFT archive.

In this chapter methods are described that were used to evaluate if three dimensional (3D) volume rendering, Infrared (IR) spectroscopy and immunohistochemistry (IHC) could be used to differentiate Epithelial Misplacement (EM) from cancer.

## 3.1 3D Volume rendering

It has been suggested within the literature that the visualisation of the continuity of the epithelium in cases of EM will be helpful in differentiating it from cancer. To enable this, volume rendering procedures were used. These processes include the acquirement of images at a suitable resolution, the transformation of these images with respect to one another and an appropriate segmentation method. In this instance, 13 EM and 9 cancer volumes were analysed.

### 3.1.1 Image Acquisition

The acquisition of images at an appropriate resolution was a very important consideration for this research project. This is because when a sample is volume rendered it first has to be sectioned into multiple slices. This causes the orientation of the tissue sections with respect to one another to be lost. Therefore, image registration algorithms have to be implemented so that this alignment can be regained.

However, before registration can take place, whole images of tissue sections have to be acquired. The usual approach of achieving this is to acquire small high resolution images and then to stitch these together[42], [119], [123–125].

#### 3.1.1.1 Scale Invariant Feature Transform (SIFT)

One way in which small overlapping image segments from the same scene can be stitched together is through the use of SIFT descriptors. In this research, the algorithm implemented was adapted from one that was published on the MATLAB file exchange [126].

In this implementation a scale space pyramid was constructed for a pair of images so that matching SIFT descriptors between them could be found. These pyramids were made up of a series of octaves that contained multiple levels, L(x,y,σ). These levels were the result of convolving an image, I(x,y), with a Gaussian Kernel G(x,y,σ) and the convolution operation computed is given by:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (6)$$

Where x and y are the spatial dimensions of an image, σ is the width of the Gaussian filter used and G is the Gaussian filter which is defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \quad (7)$$

The width of the Gaussian kernel used between levels, σ, of the octaves was separated by a constant factor, k, and from these levels a Difference of Gaussian (DoG) pyramid, D(x,y,σ), was built. The DoG pyramid is calculated by (with the nomenclature being the same as the above):

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (8)$$

It was essential for the DoG pyramid to be made as it enabled feature points from the images to be extracted [35]. These feature points corresponded to regions of maxima and minima within the DoG pyramid and were found from the implementation of a grid search operation.

The grid search compared the individual voxels of the DoG pyramid against its 26 neighbours so that keypoints could be found. Only If a feature point's value was less than or more that of its surroundings was it kept for further analysis.

This led to lots of points being found by the SIFT algorithm and these could make the algorithm too computationally expensive to implement. However, this cost can be easily reduced by removing erroneous points before the between image point matching occurs. In this instance, they were removed via the implementation of a corner detection algorithm [127], [128].

The matching of points between images was facilitated by the abstraction of pixel information that surrounded the keypoints. To ensure that the points could be optimally matched, the information used was made invariant to rotation and illumination.

This was done by translating the orientations of the pixels from the keypoint regions with respect to the most dominant found within this location. This achieved rotational invariance, as translation enables all orientations to be made relative to the dominant feature. Therefore, all orientations will be the same after this translation regardless of

the angle at which an image was acquired. For illumination invariance, weighting the intensity of the pixels surrounding a keypoint by their magnitude brought this about.

For all the retained key points, which are now known as SIFT descriptors, a 128 element feature vector was made. This feature vector contained the aforementioned orientation and intensity information that was collected for pixels that surrounded a keypoint. Computing the minimum Euclidean distance between all of the SIFT descriptors feature vectors found in each image then enabled points of correspondence between them to be defined [35], [127].

#### 3.1.1.2 Normalised Correlation

Another commonly used method to stitch images together is Normalised Correlation (NCC) and it can also be used in the linear and non linear registration of images [24], [34], [41], [42], [119].

NCC is a method by which the variance and covariance of two image datasets is compared. It will be at a maximum when they are the same. Normalised correlation is easily computed and is given by the equations presented in equation 1 and 2.

In this instance, NCC was used to refine the output of the SIFT algorithm when overlapping images taken of the same scene were stitched together. It is because the images are of the same scene that NCC can be used. This is because the overlapping regions intensities will be highly correlated.

To refine the SIFT output 100 by 100 pixel segments were defined around the matched points within images and used to define a translation that would stitch images together. Only the regions that related to SIFT points that had a NCC of above 0.95 were retained for further analysis (Figure 3.1). A translation was then defined when the distance between two consecutive SIFT points was the same.

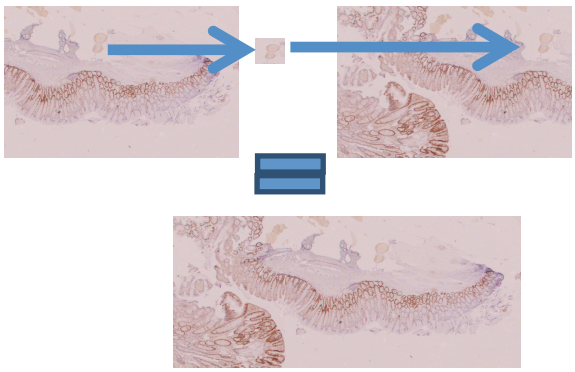


**Figure 3.1.** The process of using normalised correlation to stitch smaller images of the same scene together.

#### 3.1.1.3 RANdom SAmple And Consensus RANSAC

Another method of refining SIFT is through the implementation of the RANdom SAmple And Consensus algorithm (RANSAC). In this research, a MATLAB file exchange publication was used [129]. Briefly, the algorithm operated by:

- Fitting a model to an initial estimate of inliers;
- The model was fitted to all data points within the population of correspondences and new points became included in the estimate if they were of a good fit. The goodness of fit was defined by an error metric that was relative to the rest of the data;
- A new model was estimated from the potential inliers; and
- It was tested again against all data points and stopped when an error metric fell below a threshold or when it became static over a period of time.

In this research, the model fitted to the data consisted of a series of translations (*T*), rotations (**R**) and scaling factors (S). The optimised translations were then used to stitch images together. In this investigation, RANSAC was also used to refine SIFT points that were to be used in the linear registration of images.

#### 3.1.1.4 Phase Correlation

Finding the maximum phase correlation coefficient between images was also used to stitch images together. The translation found in this approach corresponded to the maximum peak found in the inverse Fourier transformed phase correlation matrices made for images[124], [130].

Due to the periodicity of the Fourier domain and the fact that there may be lots of locations that are highly correlated, multiple peaks will be present within this matrix. Therefore, a region was selected around the location of each maximum from within the original images. The NCC for these regions was then computed and the one that gave the highest coefficient was used to determine the required stitching translation [124]. This method of generating points of correspondences was also used in the non linear registration of images [41], [42].

#### 3.1.1.5 Automatic image tiling

Another popular image acquisition methodology within histology is to use automatic imaging systems. These systems allow for large high resolution images of tissue sections to be constructed easily and efficiently. The interest in these systems is obvious when viewing the literature available on the subject[131], [132].

In this investigation, two different automated machines were used to construct images representative of whole tissue sections. The first system was an Aperio T2 (San Diego, USA) scanner and was based in Leeds. It is part of the Leeds Institute of Molecular

Medicine (LIMM) and it can produce high resolution images like the one found in Figure 2.5. The specific details for this machine can be found in a recent publication by this group [42] and in a recent review [132].

The other automatic imaging system used in this research was part of the Perkin Elmer Spotlight 400 Spectrometer (Massachusetts, USA). This system has a stage attached, which is initialised before use, and is able to move in an x, y and z direction. During its operation a series of small images were taken, using a six times magnification, and were placed next to one another in the order that they were acquired. This enabled full images of histology tissue sections to be captured.

### 3.1.2 Image Registration

The purpose of producing high resolution images was for their further use within 3D volume rendering algorithms. However, the raw images produced from stitching and tiling cannot be used directly because of the problems associated with tissue sectioning [24], [25], [133]. These problems include the damage to tissue sections as they are being processed and the loss of orientation between them [24], [25], [27], [41], [74].

To correct for these defects the principles of image registration were used. Image registration refers to the accurate alignment of images with respect to one another. This facilitates the continuous representation of histology structures within a 3D volume.

#### 3.1.2.1 Linear Registration

In this research, the alignment that enabled features to be continuously represented was brought about via the application of linear and nonlinear transformations. In linear transformations the operations applied included translations (*T*), rotations (**R**) and scaling factors (S). In this instance, linear transformations were estimated by minimising the distance between correspondence points (section 3.1.2.1); minimising the Sum of Square Differences (SSD) [28] (section 3.1.2.1) between images and through the use of the Principal Axis Transformation (PAT) algorithm (section 3.1.2.1).

##### *Block Matching*

One way in which points of correspondence can be generated is through the block matching algorithm. It produced a similar output to SIFT (section 3.1.1.1), another correspondence point algorithm, and produced coordinates that optimally match one region of an image to another.

In this instance, block matching was carried out by comparing local blocks of a moving image against similar regions within a static image via normalised correlation[24], [34].

When the correlation coefficient was maximised a correspondence point was defined. The maximum found was the result of conducting an exhaustive search whereby a block from the moving image was compared to the entire local neighbourhood of a reference image.

This was repeated for all blocks within a moving image so that a full list of correspondence points was defined. However, these points cannot be used in the direct estimation of a transformation that will map one image to another. This is because of the presence of outliers and so a least squares optimisation was implemented to circumvent this problem. A discussion of the optimisation method used is saved for later (section 3.1.2.1) as it has applications within other methods yet to be described.

*Sum of Square Differences (SSD)*

The generation of correspondence points only had application in the registration of images that were derived from similar modalities. They had limited application within the registration of images derived from multiple modes.

This is because of the inherent differences that exist between multimodal images. Therefore, this research, registered images from different modalities by optimising an SSD metric [28]. This involved the production of binary images for each mode and this was done by using a clustering algorithm with the first principal component from Principal Component Analysis (PCA) [26].

The SSD between images was computed by calculating:

$$\mathrm{SSD} = \sum_{\mathrm{i}=1}^{\mathrm{Xdim}} \sum_{\mathrm{j}=1}^{\mathrm{Ydim}} (\mathrm{I1} - \mathrm{Tr}(\mathrm{I2}))^2 \quad (9)$$

In the above $X_{dim}$ and $Y_{dim}$ are the spatial dimensions of the images being registered, i and j are the iterated pixel positions, Tr is a transformation composed of translations (*T*), rotations (**R**) and scaling factors (S) and I1 and I2 are the images to be registered. The transformation parameters found were the result of minimising the function in equation 9 and this minimisation was brought about by implementing a simplex search method (section 3.1.2.1) [134].

*Principal Axis Transformations (PAT)*

The optimisation routine used in the estimation of transformations is often subject to the problem of local minima. This will mean that in some instances the images derived from different modalities will not be correctly registered.

The impact of local minima can be reduced via the implementation of a preliminary image transformation. A common initial deformation technique used within the literature is the Principal Axis Transformation (PAT) technique [38], [133].

PATs evaluate the first and second central moments of a binary image so that rotational and translational parameters can be defined. The translation parameter, *T*, was found from computing the image's first central moments and these are defined in equation 10.

$$Centroidx = \sum_{i=1}^{N} xb(x,y) \quad Centroidy = \sum_{i=1}^{M} yb(x,y) \ (10)$$

In the above N is the length of the rows of an image and M is the length of its columns and b is the binary representation of an image. The translation estimated by a PAT comes from the Euclidean difference calculated between the centroids of two images (i.e. the difference between centroid x and y for image 1 and image 2).

PATs also enabled the rotational elements, **R**, of the linear transformation to be defined. This came from computing the eigenvectors of the second central moments matrix made for binary images [38], and this matrix is given by:

$$Orientation = \begin{bmatrix} \sum_{i=1}^{N}\sum_{j=1}^{M} xyb(x,y) & \sum_{i=1}^{N}\sum_{j=1}^{M} x^2\, b(x,y) \\ \sum_{i=1}^{N}\sum_{j=1}^{M} y^2\, b(x,y) & \sum_{i=1}^{N}\sum_{j=1}^{M} xyb(x,y) \end{bmatrix} (11)$$

The notations of equation 11 are the same as 10. To derive the rotation required the difference between the orientations of two binary images was taken. The rotational and translational parameters estimated by the PAT were used as a starting point on which an optimisation routine improved.

*Optimisation*

The optimisation routines implemented differed between the types of images that were to be registered. For multimodal image registration, the SSD function, equation 9, was minimised in the determination of a transformation. To estimate a linear transformation from the output of points of correspondence algorithms the L1 estimator was minimised. The L1 estimator is defined by:

$$\left(\hat{\mathbf{R}}, \hat{t}\right) = \arg\min_{R,t} \sum p(\|r_k\|) \ (12)$$

Where

$$p(x) = |x| \ (13)$$

In the above, **R** and *t* hat are the estimated rotational and translation parameters and $p(r_k)$ is the absolute distance between points. This enabled accurate translation (*T*) and rotational (**R**), and in some cases scale (S) parameters, elements of the linear registration to be defined. The optimisation routine used was part of Matlab's optimisation toolbox and this uses the simplex search method in its operation [134].

#### 3.1.2.2 Non linear Image Registration

A review of the literature suggests that optimised linear registrations are not always suitable because of the defects which arise within tissue samples when they are sectioned [24–26]. To circumvent these problems, non linear image transformations were applied as they eliminate linear registration errors [40]. However, it should be noted that linear registrations provide good starting points that non linear transformations can improve upon.

There are many non linear algorithms to choose from but "as rigid as possible" schemes were sought[31], [40], [41]. This is because these methods maintain the original integrity of images and do not allow them to become over warped. This is in contrast to other methods such as Thin Plate Splines (TPS) which can easily allow images to become over deformed [39].

*Moving Least Squares (MLS)*

An example of an "as rigid as possible" technique is the Moving Least Squares (MLS) algorithm. It warps images by using local transformations that minimise the distance between predefined points of correspondence. In this instance, a program found on the Matlab file exchange was adapted to facilitate transformation by the MLS [135].

The warping is facilitated by the fitting of a grid to an image and then moving the vertices of this grid so that points of correspondence between images become optimally mapped onto one another. The algorithm was kept rigid by weighting the images correspondence points by their distance away from the moving grids vertices. This meant that only those points close to the vertices had an affect on their movement and prevented outliers from having a global effect on the overall registration.

The MLS algorithm can be adapted so that for local regions linear, affine and similarity transformations can be calculated. However, in this instance, local linear transformations were computed as it enabled greater fidelity between the post and pre transformed image to be achieved. The translation (*T*), rotation (**R**) and scale (S) parameters of the local linear transformations came from the least squares minimum of equation 14:

$$f(p,q) = min \sum_{i}^{N} w_i |I_v(p_i) - q_i|^2 \ (14)$$

In the above *p* and *q* are the correspondence points that were present within the images to be registered, Iv is the rigid transformation that was optimised by a closed set of equations and w is the weighting coefficient applied to the $i^{th}$ point. The weighting was calculated by computing:

$$w_i = \frac{1}{|p_i - v|^{2\alpha}} \ (15)$$

Where v is the vertex under consideration and alpha can be regarded as being a smoothing constraint and has to be greater than or equal to one. The local rigid transformation used was derived from the computation of:

$$Iv(p_i) = p_i \mathbf{M} + T \ (16)$$

Where in the above **M** is the rotation matrix and *T* is some translation parameter in the x and y direction. Equation 16 is evaluated for all points within an image that was to be deformed and was used with equation 14 to bring about the nonlinear warping of images with respect to one another [31].

**M** and *T* are found from functions that are dependent on the type of transformation that is sought. For instance, the equations used in the rigid implementation of MLS are different to those where an affine deformation is computed. In this instance, only the rigid implementation of MLS was used and the calculations required for this transformation are outlined in a paper by Schaefer et al [31].

The implementation of this algorithm in the generation of 3D volumes involved sequentially defining images within a stack as moving and reference images. Once a moving image had been deformed it was set as the new reference image and the next image in the sequence was warped with respect to it. This process was repeated until all of the images within a volume had been transformed to look like its predecessor.

### *B Spline grid deformations*

The computational complexity of the MLS algorithm increases with the number of vertices and the size of the image. This cost was often too large for its implementation using Matlab running on a standard desktop.

Alternatives to this technique include B Splines [29] and were used in this instance to deform histology images to look like the same and other modalities. B spline grids were used in the algorithm implemented by the LIMM. This involved minimising the Euclidean distance between points of correspondence[13], [41].

The correspondence points they generated were a result of maximising the phase correlation coefficient (section 3.1.1.4) between images within a stack [41], [130]. In this instance, registration was achieved using a subsampling methodology. Subsampling refers to the fact that the images were reduced in size and a coarse deformation refers to the use of a small B Spline grid.

Once an optimal registration had been achieved at a particular resolution and grid size, the subsampling was reduced and the number of splines was enlarged. This was repeated until the maximal resolution of the images and the algorithm's maximum number of splines, which was 8 by 8, had been reached.

However, as previously mentioned this method only has application within images that are derived from the same modality. Therefore, when multimodal images were registered using B splines a SSD metric was optimised. This required that a Boolean transformation was applied to these images and this was achieved by either using Principal Component Analysis (PCA) (section 3.1.3.1) with the k means algorithm (section 3.1.3.6 & 3.1.3.7) or by using an edge filter [136].

The binary conversion of images was necessary as using B Splines with a SSD metric to deform images will be affected by differences in intensity between images[29]. Registration is brought about a B Spline grid because the movement of its verticies enable the optimum of the SSD to be found [29], [120]. This corresponded to the optimal warping of one image so that it looked like another.

When multimodal images were registered a subsampling methodology was used that initially involved a coarse transformation with a subsampled image, in a similar fashion as before. However, the final deformation grid used with this algorithm was 16 by 16.

Cytokeratin (CK) and Haematoxylin and Eosin (H&E) images were registered in this way so that epithelial structures could be segmented (section 3.3.1.3). When registering images of this sort, edge filters were used on the image's average colour channels so that similar features between them became more obvious.

However, missing tissue sections can be a problem for the above procedure. It is estimated that up to 30% of a tissue block can be lost during processing [27] and will mean that some images will not have an alternatively stained counterpart. To enable images for these sections to be found, the CK images were selected and their H&E equivalents were deformed with respect to them.

In this instance, the CK and H&E images for a sample were independently registered by the LIMM. This is because it is difficult to establish correspondence points between multimodal images. For each CK image, the nine images before and after its position were selected from the H&E volumes, where possible, and those that minimised a coarse transformations SSD metric to the greatest extent were selected as corresponding.

These were then used with more refined registration parameters and allowed for the automated selection and optimal registration of the estimated correspondence pairs. However, on occasion, the images selected as corresponding did not overlap with a great degree of accuracy upon more refined registration. This could have been due to tears within images and so these image pairs were removed from the fused datasets.

### 3.1.3 Image segmentation

The purpose of image registration was to construct accurate 3D volumes and to bring about intermodal structural correlation. However, the interesting information within the warped datasets is not explicitly available.

To derive this, image segmentation algorithms were implemented. Within the literature there are a plethora of methods to choose from and the ones implemented in this research are documented below.

#### 3.1.3.1 Principal Component Analysis (PCA)

A well-known segmentation technique used by this research was the Principal Component Analysis (PCA) algorithm. It is a fundamental statistical technique that looks for patterns within large datasets and was used here as a starting point for the segmentation of a variety of image features.

The patterns identified by PCA enabled a dataset's total variation to be explained by a smaller number of principal components. The first principal component always describes the largest amount of variation within a data set and as the number of principal components increases their significance decreases. The matrix algebra associated with PCA is given below:

$$\mathbf{X} = \mathbf{T} \cdot \mathbf{P} + \mathbf{E} \quad (17)$$

In equation 17 **X** is the original data matrix obtained consisting of *M* rows and *N* columns which can be decomposed into a scores matrix, **T**, consisting of *M* rows and *A* columns, where *A* represents the number of principal components, multiplied by a loadings matrix, **P**, consisting of *A* rows and *N* columns plus some residual error matrix, **E**, because only *A* components are retained.

Equation 17 is represented pictorially within Figure 3.2. It demonstrates how the original data matrix, **X**, can be decomposed into a number of scores, **T**, and loadings, **P**. This figure also demonstrates the composition of the data that was used with PCA. In this research, the rows of the matrix used within PCA were made up of the investigated samples and the columns contained their characteristic data values.

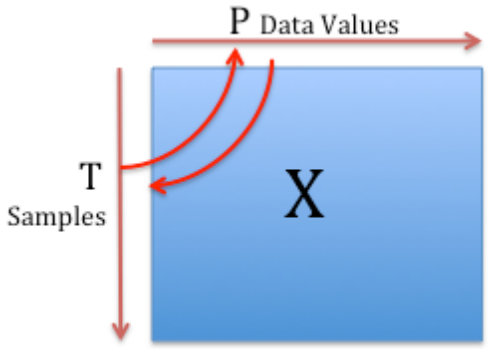



**Figure 3.2.** A pictorial example of equation 17.

For example, when using PCA with spectroscopy data, the row numbers of the matrix represents the individual spectra and the columns contained the samples Infrared (IR) absorption intensities at monotonically increasing wavenumbers. When using this matrix design, the PCA model will describe the multivariate variation that exists between samples.

Other than the spectroscopic datasets recorded by this research, PCA was also used to analyse IR chemical maps and in the exploration of features contained within Red, Blue and Green (RBG) images. When considering image datasets, the rows of the matrix used with PCA represented the pixel numbers of the reshaped images and the columns contained their intensity values. This required the images to be reshaped and a pictorial representation of this is displayed within Figure 3.3.

In this instance, images were analysed by reshaping the N by M rows of an image at a particular channel, which for instance could be one of the RBG colours of an image, into one large column vector. This was necessary because of the way in which the covariance matrix associated with PCA was calculated. The same process happened to the channel next to this and this formed the data matrices second column entry. This continued until all the channels associated with the original matrix had been formatted in this way, i.e. all image pixels formed the rows of the reshaped matrix and the columns contained the data values associated with them within particular channels (Figure 3.3).

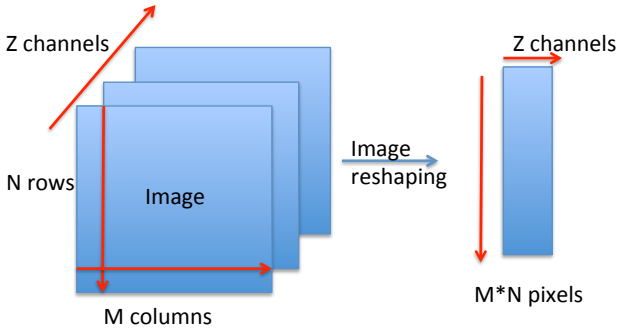



**Figure 3.3.** The image reshaping applies to Red, Blue and Green (RBG) images and Infrared (IR) chemical maps.

After decomposition by PCA, the scores of the model were reshaped back into an N by M matrix. False colouring was then applied to this image and facilitated the production of a heatmap that highlighted the pixels most strongly associated with the trend identified by the PCA score of interest. If PCA was conducted on image datasets within the subsequent chapters then the matrix design presented above was used.

### 3.1.3.2 Colour deconvolution

Alternatives to using PCA in the separation of staining colours included colour deconvolution algorithms. In this research the equations as laid out by Berger et al were used and this implementation required desired, undesired and background colours to be defined. These entities were selected from the PCA scores of a test image.

Once defined, they were used to reshape the image's RBG basis vectors into a new colour space where the undesired stain was removed. The colour space transformation, undesired colour component removal and conversion back to the original RBG space is given by:

$$\vec{c}' = \frac{\mathbf{M}\vec{c} + (d_3n_2p_1 - d_2n_3p_1 - d_3n_1p_2 - d_1n_3p_2 + d_2n_1p_3 - d_2n_1p_3 - d_1n_2p_3)\vec{u}}{(d_3n_2u_1 - d_2n_3u_1 - d_3n_1u_2 + d_1n_3u_2 + d_2n_1u_3 - d_1n_2u_3 - d_1n_2u_3)} \quad (18)$$

In equation 18, *d,n,u* and *p* stands for desired, undesired, their cross product normal and the background colour components respectively with the subscripts 1,2 and 3 standing for red, green and blue and the matrix **M** is defined as:

$$\mathbf{M} = \begin{bmatrix} d_2n_1u_3 - d_1n_2u_3 - d_3n_1u_2 + d_1n_3u_2 & (d_3n_1 - d_1n_3)u_1 & (d_1n_2 - d_2n_1)u_1 \\ (d_2n_3 - d_3n_2)u_2 & d_2n_1u_3 - d_1n_2u_3 + d_3n_3u_1 & (d_1n_2 - d_2n_1)u_2 \\ (d_2n_3 - d_3n_2)u_3 & (d_3n_1 - d_1n_3)u_3 & d_3n_2u_1 - d_2n_3u_1 - d_3n_1u_2 + d_1n_3u_2 \end{bmatrix} \quad (19)$$

The nomenclature and the subscripts in the above have the same meaning as those in equation 18 [59].

#### 3.1.3.3 Active Contour Without Edges

However, using only colour intensity in the optimal segmentation of images can be difficult to achieve. This is because of the problem of over, under and non specific staining. Therefore, textural segmentation methods were used by this research because they are somewhat invariant to the intensity of staining.

In this research, the benefit of textural segmentation was investigated by using Active Contours. There are many different varieties of active contours and each has its associated advantages and disadvantages. In this instance, active contours were grown via the principles of contraction and dilation [63].

Contour evolution by contraction and dilation was facilitated by following the rules outlined below:

- The algorithm begun by computing a "force" for the edge pixels that were found on an initial contour boundary. The force is defined as:

  $$F = ((I(x) - v)^2 - (I(x) - u)^2)(20)$$

  In the above, I(x) is the intensity of a pixel located on the boundary of a contour and u and v represent the average intensities inside and outside of a contour respectively;
- If this force is negative for a pixel directly in front of the edge of a contour then this point was given a value of 1 and if it is positive it was set to -1 (which indicated that it was inside the contour);
- The contour was evolved in this way until equation 20 was below a threshold value or until a maximum number of iterations had been achieved [62], [63];
- At each iteration, a clean up operation was also implemented and prevented odd shaped contours from forming. The clean up involved filling in holes which were within the interior of the evolving contour. These holes arose from errors caused by the evolution process.

Active contours without edges require seed points in order for the optimal segmentation of desired features to take place. In this instance, the seeds were a result of applying a colour deconvolution algorithm (section 3.1.3.2) to the images contained within a volume. The submucosal staining colours were chosen as the desired component because this caused the epithelial regions within the final deconvolved image to become white, when using Haematoxylin and Eosin (H&E) images. These white regions were then selected as epithelial specific seed points and enabled the active contours to grow within features of interest.

#### 3.1.3.4 Consensus Principal Component Analysis (CPCA) Algorithm

The growth of contours can be computationally intensive and so other segmentation methods were sought. A novel way to achieve the accurate segmentation of epithelial components from within images was through the process of combining different modalities together. This follows on from the ideology of the literature which has previously combined image datasets so that the biology of particular pathologies can be characterised [58].

However, there is no evidence within the literature of using segmentation algorithms on combined images that are derived from complementary modalities. In this instance, segmentation was facilitated by the implementation of the Consensus Principal Component Analysis (CPCA) algorithm.

CPCA is analogous to performing PCA on two independent measurements made for a system and then carrying out a further round of PCA on the initial scores. In this way, the correlated trends between the two can be quantified and allows for features of interest to be identified.

Although, in reality, CPCA [137] uses weighting coefficients with the original data so that the optimal matching between their variances is achieved. Treating each data set as a block and using PCA iterations with appropriate weights in the determination of similar trends brought this about. The weights determine how significant a block is with regards to the observed trends.

In this instance, CPCA allowed the epithelium to be segmented from within fused H&E and CK. The protocol required to derive CK sections is detailed in Table 3.1. These reagents were chosen because together they are very specific for the epithelium. This is because the MNF116 CK (DAKO, Cambridge, UK) antibody specifically stains the epithelium of a tissue section brown. However, it will also erroneously stain blood vessels the same colour. Therefore, both the epithelium and blood vessels would be preserved if only the CK staining colour was used in the segmentation process.

The blood vessel staining by the CK antibody can be removed if its consecutive tissue section was stained with H&E. This is because the H&E reagents will stain the vessels of this section red. The non linear registration (section 3.1.2.2) of images of the contiguous sections with respect to one another means that the red staining of the H&E and the brown of the CK for the blood vessels will overlap. Registration also enabled the purple staining for the epithelium within the H&E image to be aligned with the brown staining for this feature in the CK image

CPCA easily identified these trends and produced a score that only represented where the epithelium was located. To ensure that this analysis could be conducted on all samples, they were all alternatively stained with CK and H&E and images of these were converted to the L*a*b colour space. These volumes were independently registered with respect to one another and so a correspondence algorithm was used to identify image pairs, as previously discussed, before application of CPCA.

Further details on how CPCA was applied to the co-registered CK and H&E images found within this researches volumes is presented within Figure 3.4.

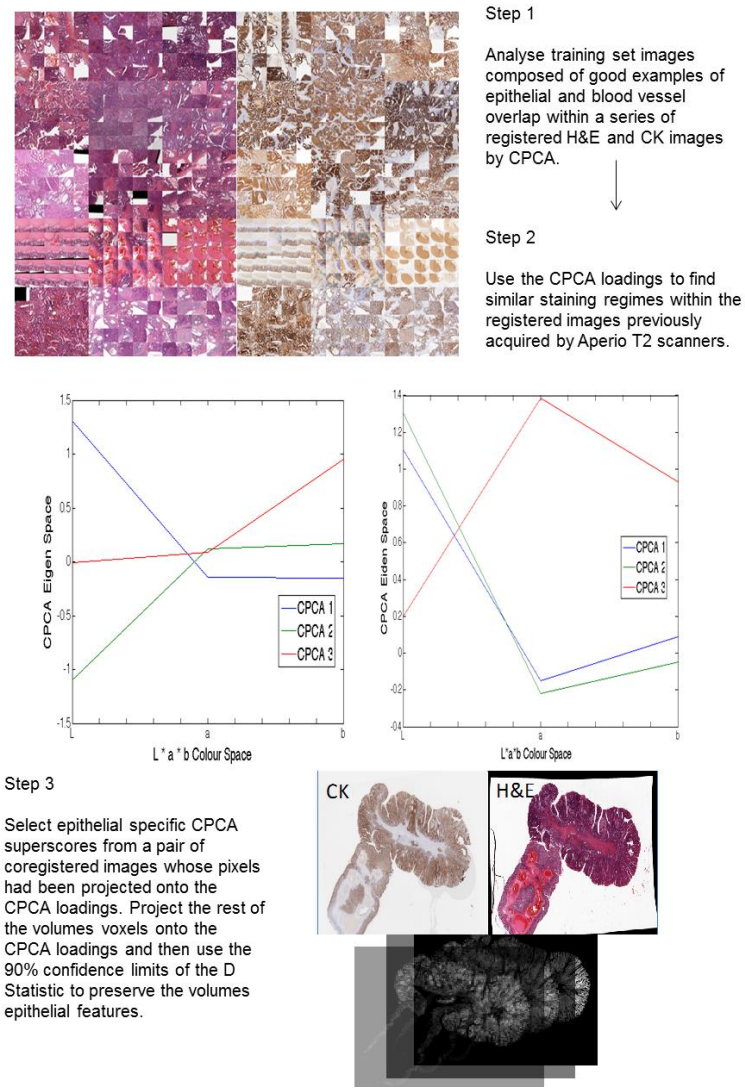



**Figure 3.4.** A schematic representation of how CPCA was utilised for the purposes of epithelial segmentation.

In this instance, a training set of images, Figure 3.4, was constructed by selecting regions from the intermodally registered image volumes that were thought to be good

examples of epithelial and blood vessel overlap. This training set of images was reshaped into a column vector, Figure 3.4, and its CPCA loadings were calculated.

The CPCA loadings derived from the training images were then applied to each of the co-registered images contained within a samples volume. From this volume, epithelial specific super scores were selected and used with the 90% limits of the D Statistic so that a binary volume could be made that contained the volumes epithelial features. From these Boolean volumes a measure of epithelial connectivity could be derived, which in this instance amounted to recording the volumes number of 26 3D connected components.

#### 3.1.3.5 D Statistic

The D statistic is a multidimensional t test and was used to segment the epithelium from within entire volumes. In this instance, the D statistic was applied to training and test set CPCA scores and the training set was representative of good examples of blood vessel and epithelial region overlap, Figure 4.19. The loads and weights calculated from these were used to predict new CPCA scores for other stacks and epithelial specific scores were selected from these. This enabled the D Statistic's 90% confidence limits to be calculated.

The D Statistic uses the Mahalanobis distance between the training, *T*, and the test set scores, $\hat{\boldsymbol{t}}$, to reject those that are unlike the training set. The rejection thresholds are determined by confidence limits set by the F distribution [138]. The D statistic was calculated by using an inhouse built script and is computed via the following equation:

$$D = \hat{t}\left(\frac{T'T}{I-1}\right)^{-1}\hat{t} \quad (21)$$

In the above equation I stands for the number of samples within the training group, *T*, and, $\hat{t}$, stands for the test group. Both *T* and $\hat{t}$ were mean centred. The D statistic also had other areas of application in this research, as it was also used in the segmentation of mucin from within IR maps.

#### 3.1.3.6 Clustering Algorithms

Other ways of segmenting images were also investigated. For example, K means clustering was used to automatically select tissue components from within PCA scores that were computed for IR and standard histology images. This facilitated the Boolean transformation of images and allowed intermodal registration to occur. The K means algorithm used in this research came from an inhouse built Matlab script. Although, a D Statistic approach could have been taken, this option provided a simpler alternative.

The K means algorithm operates through two stages that are known as the assignment and update phases. In the assignment step an observation, the intensities of pixels in this case, is classified as a particular group based upon its Euclidean distance away from the mean of clusters. In the initial implementation of the algorithm the points used to represent the mean are randomly generated.

Class membership for all points is computed and enables the update phase to calculate new means for the clusters. This process repeats until there is no change in the value of the means calculated or until a maximum number of iterations had been reached [26].

Another clustering algorithm investigated is known as Fuzzy C Means (FCM). In this approach, pixels could be part of all clusters to a certain degree. For example, a data point maybe fully part of cluster one or it could be part of cluster one and two [65]. In this algorithm, the weighted distance of the data points away from the mean of clusters was used to determine the degree of class membership. The weights were iteratively updated until there was no change in the calculated means, relative to a threshold. The weights are determined from previous iterations and are equal to the inverse of the data points previous distances away from the clusters means.

This algorithm can also be adapted so that the neighbours of pixels within images have an influence on class membership. This was facilitated by adding an additional term to the update stage of the FCM that corresponded to the distance of the pixel's neighbours away from a given cluster. It was thought that this would aid in the removal of erroneously stained pixels.

## 3.2 Immunohistochemistry (IHC)

Volume rendering allows for the epithelial continuity of tissue samples to be investigated but it is only one way in which the literature suggests that differentiation can be brought about. For example, it has been noted that the profiling of tissue samples by immunohistochemistry (IHC) is useful as they enabled classificatory biomarkers to be found. Therefore, this investigation conducted independent analysis of the antibodies identified in the literature so that their claims could be scrutinized.

This involved analysing a pathologist's assessment of staining strength and the results were used further with the Mann Whitney U statistic. This facilitated the determination of significant differences and the methods used in this investigation are presented below.

### 3.2.1 IHC staining procedures

The laboratory processing of IHC stained tissue sections was conducted by members of staff working within the pathology department of Cheltenham Hospital (Gloucestershire Hospitals NHS Foundation Trust (GHNFT), UK).

The antibodies they used in the staining of tissue sections were Ki67, Matrix Metalloproteinase 1 (MMP 1), Collagen IV, E-Cadherin and p53 [11–13]. These antibodies were used to stain ten cases of EM and cancer respectively. However, it should be noted, that one of the EM slides stained for Collagen IV was destroyed prior to analysis.

Analysis of IHC stained tissue sections required that the staining be of the highest quality. Therefore, a DAKO (Cambridge, UK) Envision Flex High pH buffer system was used, except in the case of Collagen IV, to ensure this.

The antibodies were diluted from stock, Table 3.1, and left to hybridise with tissue sections for an amount of time as specified by the manufacturer. They were also counterstained with Mayers haematoxylin (BDH, Leicester, UK) so that a contrast between stained and non-stained regions could be seen.

**Table 3.1.** The antibodies used by this research and the dilutions made from stock to stain the tissue sections. a= (DAKO, Cambridge, UK) & b=( Neomarkers, Michigan, USA).

| Antibody | Dilution | Manufacturer | Sample Size EM | Sample Size Cancer |
|---|---|---|---|---|
| **Ki67** | 1/100 | DAKO[a] | 10 | 10 |
| **p53** | 1/500 | DAKO[a] | 10 | 10 |
| **MMP-1** | 1/50 | Neomarkers[b] | 10 | 10 |
| **Collagen IV** | 1/50 | Neomarkers[b] | 9 | 10 |
| **E-Cadherin** | 1/100 | DAKO[a] | 10 | 10 |
| **MNF116 Pan Cytokeratin** | 1/50 | DAKO[a] | N/A | N/A |

### 3.2.2 Histological Grading of tissue sections

The grading of tissue sections enabled the abundance of proteins to be qualitatively analysed. This analysis was conducted for all IHC antibodies except in the case of Cytokeratin, as this was not evaluated for the purposes of differentiating disease states.

In this instance, the intensity and extent of staining in the submucosal epithelial islands was compared to that of the samples surfaces. In the case of E-Cadherin, p53, Ki67 and MMP 1 differences in staining intensity between these regions were noted. For these stains, they were classified as either having no staining or they were classified as being

mildly, moderately or intensely stained. In this analysis, if they had no stain they were given a value of zero and for the other groups a value of between one and three was given respectively.

For Collagen IV, the continuity of staining within the basement membranes of the submucosal glandular and surface tissue was analysed. If there was no staining then the samples were given a value of one and they were given a value between two and three if the staining was discontinuous or continuous.

### 3.2.3 The Mann-Whitney U Statistic

In this instance, the Mann-Whitney U test was used to determine if any significant differences in staining existed between the EM and cancer cases. The Mann-Whitney U test works by comparing the sum of ranks between distributions [87] and enables the analysis of ordinal variables.

To carry out this test, two data sets are ranked by their order of magnitude and are individually given a score of +1 every time they outrank the other distribution. The highest rank sum was then used with the following formulae so that the differences between the two distributions can be compared:

$$U_k = n_1 n_2 + \frac{n_k(n_k + 1)}{2} - R_k \quad (22)$$

Where U is the Mann-Whitney U statistic, n is the size of the dataset, R is the sum of the ranks for the $k^{th}$ data set and k is the dataset with the largest rank. A critical value for U was determined from the normal distribution and the null hypothesis was rejected if it was more than this value.

In this instance, the test was applied to continuity metrics derived from 3D volumes and to the IHC tissue sections staining intensity results obtained from a senior pathologist. In the former, the null hypothesis was that the staining intensities between the two different pathologies were the same. In the latter, the null hypothesis was that there was no difference in connectivity within 3D volumes made for cancer and EM cases. In this investigation, not only where univariate tests conducted on the IHC scores but multivariate analysis, in the form of Hierarchical Cluster Analysis (HCA), was also carried out.

### 3.2.4 Hierarchical Cluster Analysis (HCA)

Hierarchical Cluster Analysis (HCA) group's data together using the distance properties of the objects found within the feature space made from the IHC scores. Classification is performed using an agglomerative approach, sometimes referred to as bottom up. The algorithm starts by considering all elements, which in this case were the pathology samples, of the feature space.

This algorithm groups points together in an iterative pairwise fashion. Initially a distance metric, which in this instance was Euclidean, was used in the pairing together of pathologies with similar IHC scores. In the next iteration, the shortest distances between the clusters found in the previous step were used to form new groups. This process continues until n-1 iterations have been completed, where n is the number of samples, and at this point all elements of the feature space will have been given the same class label.

## 3.3 Fourier Transform Infrared (FTIR) Spectroscopy

The drawback of using IHC antibodies is that they can only stain for the presence of single molecular components within tissue sections. FTIR provides an alternative to this limitation as it probes the vibrational modes of a variety of functional groups. Therefore, FTIR can be used to visualise more than one molecular configuration within a tissue sample [86], [98], [139].

FTIR spectroscopy is not limited to just the visualisation of molecular information. It can also be effectively used in the classification of pathology groups when used with multivariate statistics [48], [50], [51], [79], [90], [92], [122].

In this instance, FTIR spectroscopy was used in a variety of ways but all with the aim of differentiating EM from cancer, Figure 3.5. In this study, 11 cases of EM and cancer were analysed. Although, it should be noted, that the multivariate model only included spectra from nine samples of cancer. In an additional study, the mucin spectra from four cases of EM, non neoplastic tissue and cancer were investigated.

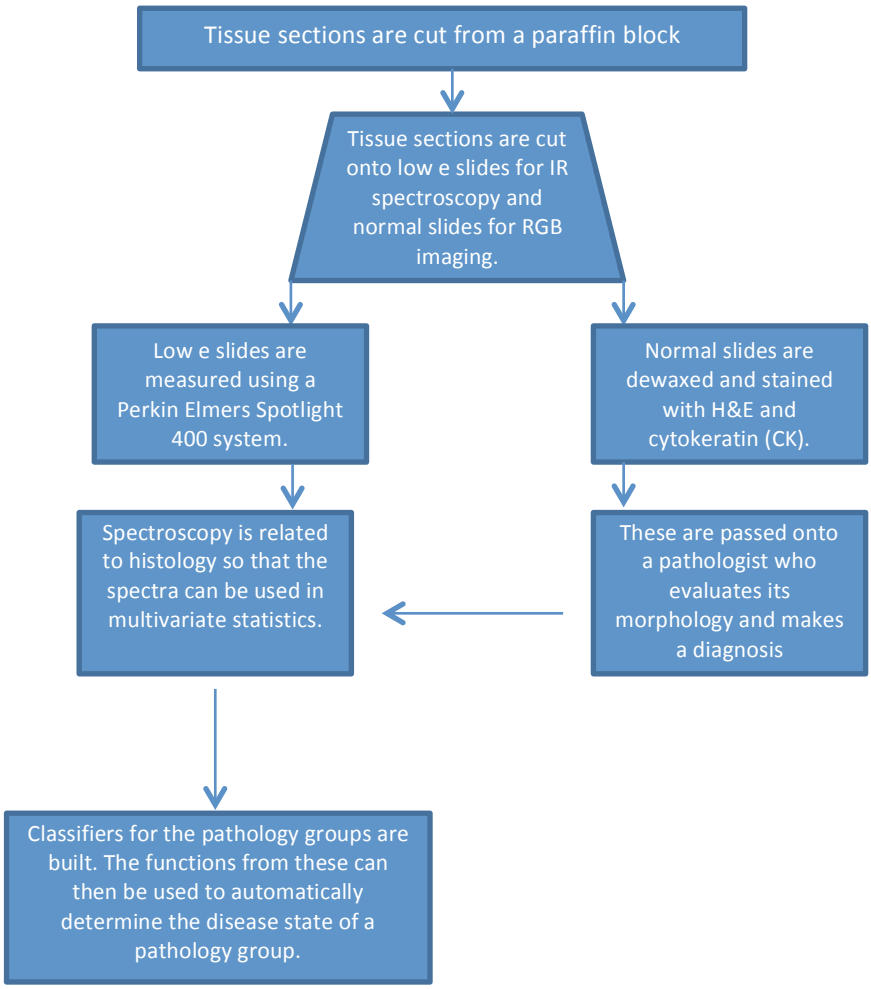



**Figure 3.5**. A flow chart describing how FTIR spectroscopy was used in the disease classification process.

In Figure 3.5, the methods used to correlate histology with spectroscopy are highlighted. It involved using a pathologist's expert assessment in the determination of a pathologies disease state and then using the spectra derived from these within multivariate models. This allowed this research too:

1. Locate particular histological features within unstained tissue sections so that they can be mapped with IR; and
2. Allowed for efficient classificatory models to be built [47], [51].

To enable the above to occur, the principles of image registration were utilised and the methodologies associated with this are described below.

### 3.3.1 FTIR Imaging

The correlation of histology and spectroscopy via registration was made possible because of the linear array detector of the Perkin Elmer Spotlight 400 Spectrometer.

This system moved a 16 by 16 array of Mercury-Cadmium-Telluride (MCT) detectors in a raster like fashion above a tissue sample so that a chemical image could be made[98].

#### 3.3.1.1 FTIR imaging and experimental parameters

The quality of the image produced is dependent on factors such as the spectral and spatial resolution of the measuring system. This investigation used IR in a multitude of different ways, each with slightly different experimental and instrumental parameters, and the impact of these choices is explained in more detail in section 4.0.

In the first instance, mucin was probed with IR and in this study spectra were:

- collected using a spectral and spatial resolution of 6cm$^{-1}$ and 6.25µms respectively;
- processed using a second derivative 15 point third order Savitzky-Golay filter (section 3.3.1.4), normalised and mean centred;
- paraffin corrected by deleting the main peaks relating to this compound from the IR spectrum and by using xylene; and
- acquired using 2 coadded scans.

Further studies measured individual glands from tissue sections and these spectra were:

- collected using a spectral and spatial resolution of 4cm$^{-1}$ and 25µms respectively;
- processed using a second derivative 15 point third order Savitzky-Golay filter (section 3.3.1.4), normalised and mean centred;
- paraffin corrected by deleting the main peaks relating to this compound from the IR spectrum; and
- acquired using 124 coadded scans.

In the last experiment, whole tissue sections were acquired using IR and these spectra were:

- collected using a spectral and spatial resolution of 6cm$^{-1}$ and 25µms respectively;
- processed using a second derivative 15 point third order Savitzky-Golay filter (section 3.3.1.4), normalised and mean centred;
- paraffin corrected by deleting the main peaks relating to this compound from the IR spectrum; and
- acquired using 2 coadded scans.

The vast majority of techniques used in this investigation have been previously explained, such as PCA (section 3.1.3.1), PATs (section 3.1.2.1) [38], and the CPCA algorithm (section 3.1.3.4). What follows next is an explanation of the remaining techniques utilised by this project.

#### 3.3.1.2 Intermodal Image registration

In this investigation, histology specific spectra were used because this aids classification. To abstract histology specific spectra an image registration protocol was developed which combined images from different modalities together.

This involved using registration techniques that were both linear and non linear in nature. For the linear transformations, the image's PCA (section 3.1.3.1) scores were calculated and used further with k-means clustering (section 3.1.3.6). In this instance, the first principal component was chosen because it always explained the variation between an image's tissue foreground and background.

Manual selection of the appropriate cluster allowed binary images to be formed that when used further with a PAT facilitated linear registration. This was followed by optimising a Sum of Squares Difference (SSD) metric.

Although, it should be noted that the well defined problem of tissue deformation prevented a reliance on linear transformations [25], [27]. It was for this reason that a B Spline grid non linear registration (section 3.1.2.2) scheme was used in order to optimally warp one image to look like another [29], [40], [120]. The intermodal registration scheme implemented here was the same as that found within the segmentation section (section 3.1.3).

#### 3.3.1.3. Image segmentation

Mapping and registration allowed for precise histological and spectroscopic relationships to be defined. However, the features of interest contained within these still needed to be derived and this was facilitated by the use of appropriate segmentation algorithms. In this instance, mucin and epithelial targets were probed with IR.

##### *Mucin segmentation*

For Mucin, segmentation was achieved by using the PCA loads of a target image, where the mucin was identifiable within its scores, with the acquired IR chemical maps. Further to this, a training set of mucin scores were used with the 95% confidence limits of the D Statistic and this enabled mucin specific spectra to be derived that could be used with multivariate statistics (section 3.3.2).

*Epithelial segmentation*

For epithelial spectra, a similar methodology as the mucin study was also implemented along with the segmentation procedure presented in section 3.1.3.4. Whereby, H&E and CK images were registered and used with the CPCA algorithm [137]. These registered images were then further transformed with respect to their IR chemical map counterparts.

The H&E, CK and IR tissue sections imaged were consecutively cut from sample blocks and this ensured that their individual morphology was similar. This allowed for one image, where a feature could be easily segmented, to be used in the exploration of another. As a result, the identified epithelial regions could be selected so that its associated spectra used further within a multivariate model (section 3.3.2).

Another procedure that enabled the collection of epithelial specific spectra was a digital staining technique (Figure 3.6). In this method, the PCA scores of registered standardly stained (H&E/E-Cadherin) and IR images were computed. In this instance, the same intermodal registration procedure as described in the segmentation section was used.

PCA was conducted so that the variance between the two images could be related and this was facilitated by a linear regression. This enabled a series of coefficients to be derived that produced functions that would digitally stain images, Appendix D.

From these images, PCA scores were computed and they were manually thresholded so that epithelial specific spectra could be derived. In this instance, the epithelial spectra that were selected related to regions that had been previously identified by a pathologist. In Appendix B, the selected regions are highlighted by blue boxes and these were used with the digitally stained images, Appendix D, in the derivation of spectra representative of a particular pathology.

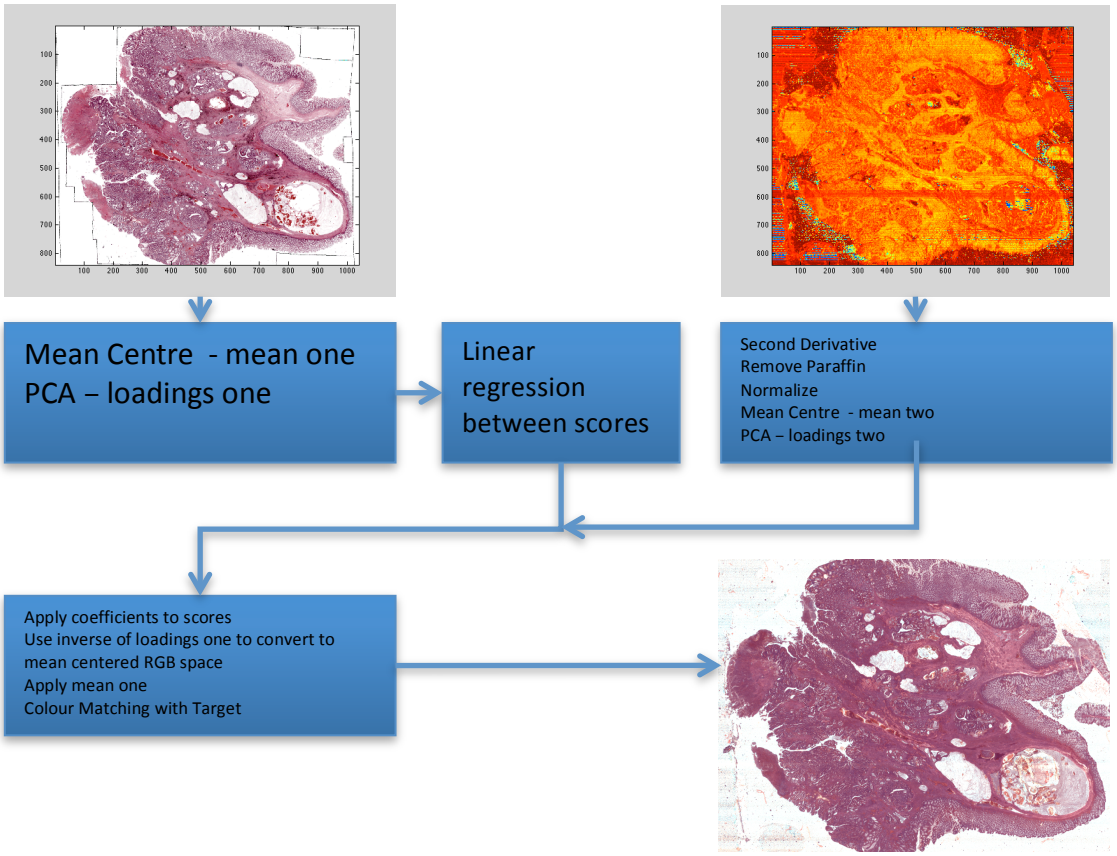



**Figure 3.6.** A schematic representation of how the Infrared (IR) Principal Component Analysis (PCA) scores were related to those of its corresponding standardly stained section by linear regression. It should be noted here that the IR map representation is a PCA score image where only the carbohydrate region of the spectrum was analyzed. Although, the regressive model built used all of the recorded spectra, except for the paraffin peaks.

*Colour balancing digitally stained images*

However, the small number of coadded scans used sometimes meant that poor spectra were acquired. In these regions the digital staining was dull and so a colour balancing algorithm was developed to correct for this. Within the literature this has usually been achieved by deconvolving the colour of a particular stain from within a reference image. The means and standard deviations of the reference stain colour are then matched to those of a target [33].

However, this technique can be particularly cumbersome as it requires the definition of undesired and desired staining colours. Therefore, a new way of matching the colours between images was developed and this was done so by matching the PCA scores of a reference and target.

The application of a reference image's PCA loads on a target transformed its colours using the same weights as those derived from the original reference. This allowed the

target's colours to be projected onto those of the reference and facilitated colour balancing.

This was achieved by matching the means and standard deviations of the reference and target image's scores. These were then converted back into the Red, Blue and Green (RBG) colour space by multiplying the transpose of the references loads by the target's redistributed scores. The further application of the references mean than allowed for images, like those presented in Figure 3.6, to be made.

#### 3.3.1.4 Extended Multiplicative Scatter Correction (EMSC)

For spectra that were collected using a rapid scan methodology, where only two scans were coadded, improvements to the dataset were required. The low number of scans can lead to spectral artefacts being present within the recorded data. However, through the application of the Extended Multiplicative Scatter Correction algorithm (EMSC), equations 4 and 5, these defects can be corrected [100], [107]. EMSC was also used to correct for any signal contribution that came from paraffin, the substance in which tissue samples were embedded [102], [104].

#### 3.3.1.5 Savitzky-Golay (SavGol) Second Derivative Filtering

Other issues, besides unwanted spectral contributions, include the fact that baseline artefacts affect IR spectra. One way in which the literature approximates these artefacts is by taking the spectrum's second derivative. It is well known that the second derivative will resolve overlapping peaks and make spectra relative to their instrument's baseline [85], [88], [140].

There are many ways of approximating the second derivative of IR spectra and one way is to use the finite differences between spectral points. However, this has the consequence of causing the noise of the spectra to be amplified.

Therefore, other methods of deriving the second derivative of spectra were sought. In this instance, the SavGol filter, which was part of the EigenVector (Wenatchee, USA) PLS toolbox, was used. Briefly, this method operates by fitting a polynomial to a series of spectral points that are found within a predefined window. The second derivative of the polynomial at the point on which the window was centred was then used to approximate the derivative of the raw spectra [141].

### 3.3.2 Multivariate statistics

Once the histology specific spectra were refined they were used with multivariate statistics in order for functions to be generated that would enable automatic differentiation to take place. In this research, PCA (section 3.1.3.1) was used on the recorded IR spectra so that its variance could be determined. The most discriminatory

principal components were selected, via ANOVA (section 3.3.2.1), and used further with a Linear Discriminant Analysis (LDA) classifier [50], [85], [86], [94] (section 3.3.2.2). This enabled automated functions to be produced that differentiated EM cases from cancer.

### 3.3.2.1 ANalysis Of VAriance (ANOVA)

An important part of the procedure above is that principal components are chosen that are representative of the most significant differences that exist between pathology groups. To ensure this, ANOVA was carried out on PCA scores.

ANOVA operated by calculating the within and between group variances and this facilitated the production of an f ratio. This ratio was compared against the critical limit derived from the f distribution and was used in the determination of significant differences [142], [143].

The ANOVA f ratio is calculated by:

$$Fratio = \frac{Variance\ between\ groups}{Variance\ within\ groups} (23)$$

If the f ratio exceeded the critical limit then the null hypothesis was rejected. In this instance, the null hypothesis was that no significant differences between the different pathology groups existed. Therefore, if an f value for a score exceeded the critical limit then there was significant variation between EM and cancer in this score.

### 3.3.2.2 Linear Discriminant Analysis (LDA)

These discriminatory scores were used within an LDA model so that a function was produced that would automatically differentiate EM cases from cancer. To derive this function, LDA uses the Mahalanobis distance to maximise the between and minimise the within group variation:

$$d_{ig}{}^2 = \left(x_i - \overline{x_g}\right)\mathbf{S_p}^{-1}\left(x_i - \overline{x_g}\right)' (24)$$

In the above d is the distance of the observation i away from group g and $S_p$ is the pooled covariance matrix which is defined by:

$$\mathbf{S}_p = \frac{(I-1)\mathbf{S}+(I'-1)\mathbf{S}\prime}{(I-I'-2)} \ (25)$$

In the above ***S*** and ***S'*** represents the variance – covariance matrix for a particular group and I is the number of samples in that class [51], [85], [94], [143].

### 3.3.2.3 Cross Validation

Only a small proportion of the population was being sampled in the PCA fed LDA model. Therefore, the generalisation of the model to unknown samples can be estimated by scrutinising its response to being cross validated. This produces sensitivity and specificity metrics that give information about how many times the

samples were misclassified. There are many validation schemes to choose from but a popular variant within the literature is known as Leave One Out Cross Validation (LOOCV) [51], [143].

In this procedure the spectra used within an initial model were tested by building a new model where cohorts of spectra were left out. The spectra left out of the new model related to the individual regions that had been segmented from the IR maps of whole tissue sections (section 3.3.1.3). The loads of this new model were then used to determine the class membership of the spectra that were not included in the new model.

If more than 50% of the tested spectra belonged to a particular group, then they were classified as belonging to it. This is because when glands are segmented there will be thousands of spectra associated with one epithelial region. Therefore, to determine the pathology of the sample as a whole, majority rules have to be applied.

## 3.4 Summary

The above methodologies enabled this project to investigate 3D volume rendering, IHC analysis and IR spectroscopy. There are many particularities associated with each technique and viable methodologies have been found that negate some of their limitations.

For example, multiple image stitching algorithms [123] were found that enabled whole tissue representations to be made. These provided datasets that could be used with registration algorithms. The “as rigid as possible” algorithms are thought to be better suited for histological registration [29], [42] and multiple varieties of these are described above.

The registration of these images with respect to one another is important but even more so is the meaningful representation of histological features. In this instance, the literature was searched for appropriate algorithms [63], [65], [67] and an investigation of these will be conducted. Also, a new intermodal registration and segmentation routine, that utilised CPCA [137], was developed so that comparisons between this algorithm and the current literature can be made.

Although, the literature has great belief that 3D volume rendering will facilitate the differentiation of EM from cancer [9], [12], other metrics were also sought. This included IHC staining alternatives and it has been found within the literature that p53, MMP 1, E-Cadherin, Ki67 and Collagen IV can be used to differentiate EM from cancer [12], [13]. These were tested again in this research so that the claims of the literature could be validated and the methods of analysis that will enable this are presented above.

The final technique identified by this investigation as having some classificatory potential is IR spectroscopy. It has been used previously in the classification of a variety of different disease types [48], [50], [84], [85], [87], [88], [94], [95], [101] because it probes the vibrational modes of molecules[50]. In this instance, a PCA fed LDA model was used with spectra derived from a mucin segmentation and digital staining technique (section 3.3.1.3).

In this next chapter, the results of implementing these methodologies are described.

# 4.0 Results

Epithelial Misplacement (EM) within the colon is characterised by the presence of glandular tissue within the submucosal regions of polyps [1], [8], [9], [12], [13], [144]. These islands of epithelium arise because of the continual fluxing of the colon and this will cause polyps to become damaged and bleed. As a result, the Bowel Cancer Screening Programmes (BCSP) Faecal Occult Blood (FOB) test will be positive so EM polyps are selected for pathological review.

These cases may get mistaken for cancer because upon histological processing the misplaced islands will look like they have disassociated from the surface by metastatic mechanisms. However, only in the cancer cases does the submucosal glandular tissue have any malignant potential[8], [145], [146]. Therefore, surgery is only justified in cases of adenocarcinoma.

Evidence that EM is a benign condition comes from an historic study that found that patients who exclusively contained these polyps did not go onto develop colorectal cancer [8]. Therefore, reliable diagnostic adjuncts are required to prevent this surgery. In this instance, the benefits of using three dimensional (3D) volume rendering, immunohistochemistry (IHC) and Infrared (IR) spectroscopy were evaluated and the results of this are presented below.

### 4.1 3D Volume rendering results

The purpose of the following study was to determine if the continuity of epithelial features within 3D volumes of EM and cancer could be used to bring about their differentiation. This involved the acquisition of images at a suitable resolution, their deformation with respect to one another, using registration techniques, and the application of a suitable segmentation technique.

In this instance, a novel epithelial segmentation methodology was developed that facilitated the evaluation of epithelial connectivity. The developed segmentation routine involved the co-registration of images of Heamatoxylin and Eosin (H&E) and Cytokeratin (CK) tissue sections. The number of 26 connected components were recorded from the segmented volumes and used with inference tests to determine if any significant differences between the pathologies existed.

Within the following chapter, the results of the above analysis are presented along with suggestions for future work.

### 4.1.1 Image generation and registration

The images used within the 3D volume rendering process are the fundamental components of this technique. The images acquired have to be of a suitable quality and of an appropriate resolution.

In this instance, three image acquisition techniques were tested to ensure a methodology was found that met the above objectives. Firstly, whole images of tissue sections were constructed from the stitching together of overlapping images (Figure 4.1). This stitching was achieved by using Scale Invariant Feature Transform (SIFT) descriptors and normalised correlation and these facilitated the computation of a translation that would perfectly overlay images [119], [123].

SIFT descriptors enabled a translation to be found because this technique finds correspondence points between images. However, a unique translation that would stitch the images together cannot be found from these as they contain many anomalies.

Therefore, the Normalised Correlation Coefficient (NCC) was implemented as a way of correcting for erroneous correspondences. NCC facilitated this through its use as a comparison metric. In this instance, 100 by 100 image regions that were defined around the images matching SIFT points were used with NCC. Only those points that were highly related to each other, defined as having a NCC of 0.95 and above, were retained for further analysis. The rationale behind this approach was that only the pixel intensities within similar regions would be highly correlated.

It is from these retained points that a translation was computed and this enabled the stitching together of images. The translation was defined when the Euclidean distance between two consecutive refined SIFT points was the same. The operational details of this procedure are presented in Figure 4.1 A-D. It can be seen that this approach was highly effective at stitching images together, Figure 4.1 D, and that the refinement worked very well. This is highlighted by Figure 4.1 A-C where it can be visually seen that the normalised correlation corrected matched green points are truly representative of similar regions. This is in contrast to those points that were rejected, the black circles in Figure 4.1 A-C, where it can be seen that for these points no true correspondences exist.

This approach required that the images used in the stitching process were of the highest quality. In this instance, the images that were used were directly acquired from a Leica DFC345FX (Milton Keynes, UK) histology camera microscope. Images that have been taken by this device will contain a significant amount of vignetting and therefore cannot be used directly in the image stitching process (Figure 4.2). This is because this

phenomenon will affect the correct matching of features between overlapping images [123].

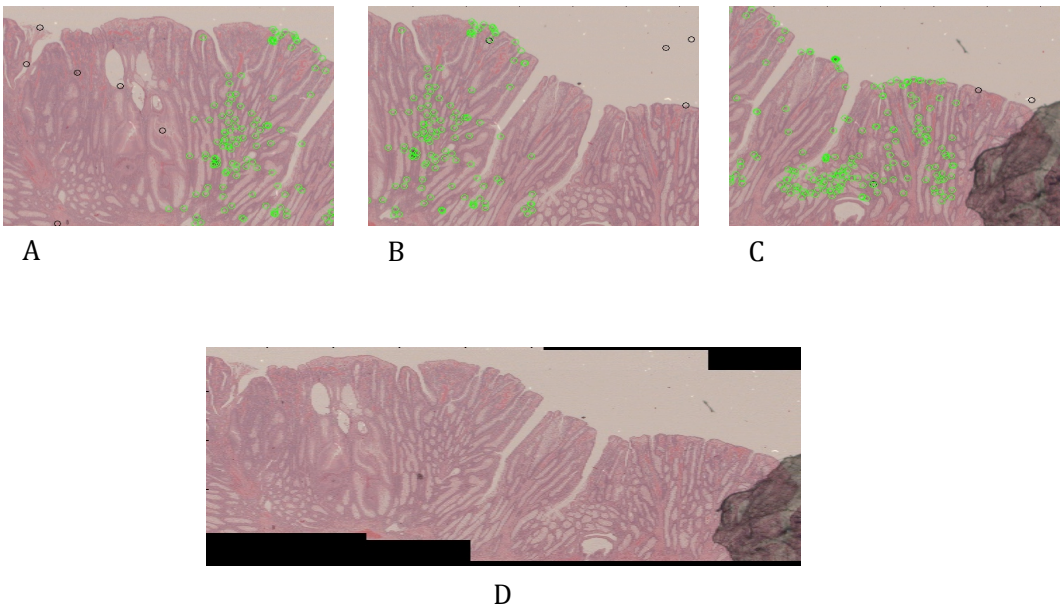



**Figure 4.1.** The stitching algorithm implemented by this research. It was used to build image representations of a scene that was larger than the field of view of the imaging equipment (D). The three images in A - C are consecutive images taken of the same scene. The green points are correspondence points that have been matched between images and have also been accepted by normalised correlation. The black points are those that have been rejected because they do not represent similar features between images.

Vignetting is the gradual loss of light intensity at the edges of images and is due to point light sources. If images of the type presented in Figure 4.2 A were stitched together then the final image would contain significant tiling anomalies. However, by application of a reference image (Figure 4.2 B), which was a piece of a slide that contained no tissue, the vignetting effects can be removed. This enabled for visually aesthetic images to be produced that could be used further within image processing algorithms (Figure 4.2 C) [25].

Generally, this sort of image stitching was accurate and computationally efficient. However, the efficiency of this approach was compromised when the size of the polyp was on the same order of magnitude as the slide on which it was mounted. This is because the computational cost of this method is proportional to the number of images taken for a tissue sample. This can be problematic, as EM cases are usually much larger than cancer. For instance, one image of a section of a case of EM took up to an hour to stitch together using this method.

A B C

**Figure 4.2.** The effect vignetting had on the output images of the histology camera microscope and how it can be corrected for. A) An image taken from the Leica histology camera microscope where the vignetting is obvious. B) The reference image that was used to correct A. C) The after effect of correcting A with the reference image B.

To reduce this cost, the images were resized but this caused the tissue section images to lose important features. Therefore, this technique is not practical for the purposes of 3D volume rendering.

Another drawback of this method was that it could not handle large areas of homogeneity, which are inevitable within histology images. For these regions the NCC will be high and so SIFT points that are not representative of the same feature will be preserved. An example of this complication can be found in Figure 4.3. It can be seen in this figure that the gland circled by the blue box contains many similar features and this led to the erroneous stitching together of images.

Therefore, other image stitching methods were sought, such as joining overlapping images together by maximising their phase correlation coefficient [124]. Although, this approach was again generally accurate, it was still very computationally expensive for large images. An example of an image produced using this mosaicking method is presented in Figure 4.4. For this one image, it took half an hour to stitch the smaller segments together. This was problematic as there were 60 tissue sections for this case and this would have meant that around 30 hours would have been spent on the construction of its images.

Therefore, automated imaging systems were investigated. It was thought that they would improve the aforementioned computational costs, the length of time required to generate images and reduce the erroneous nature of the previous methods. These systems simply involve using a motorised stage to move a sample under a camera. At each iteration of the stage an image is acquired. This allows for the individual images

to be placed next to each other in the order they were captured and this facilitated the production of large scale images (Figure 4.5 A&B) [42].

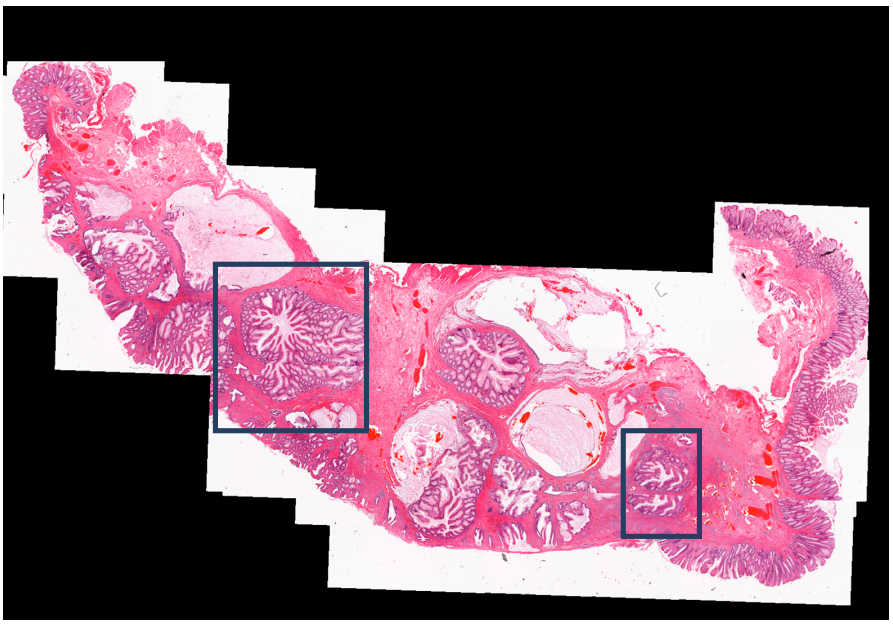


**Figure 4.3.** An example of the limitations of the implemented image stitching algorithm. Similar features that exist in the overlapping images have been identified by SIFT. Since these regions share similar characteristics they also maximised the NCC. However, they are not representative of true points of correspondence and so errors will accumulate in the image stitching process.

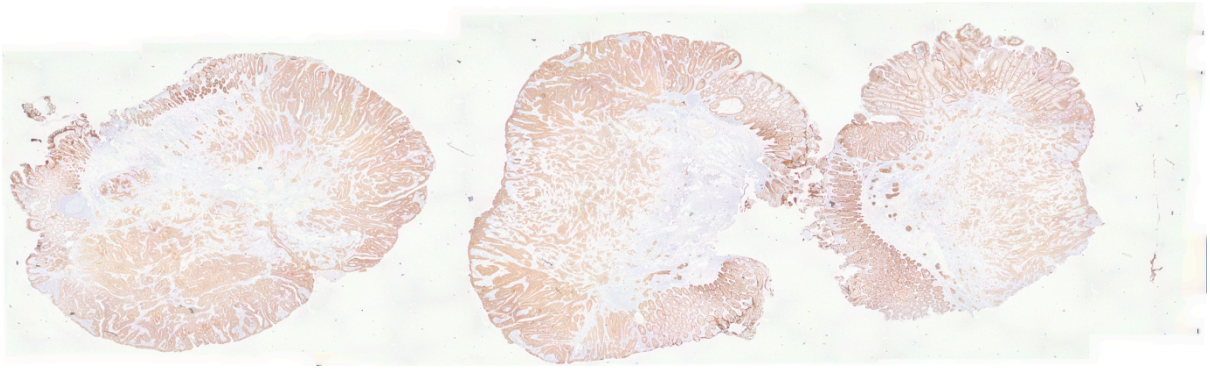
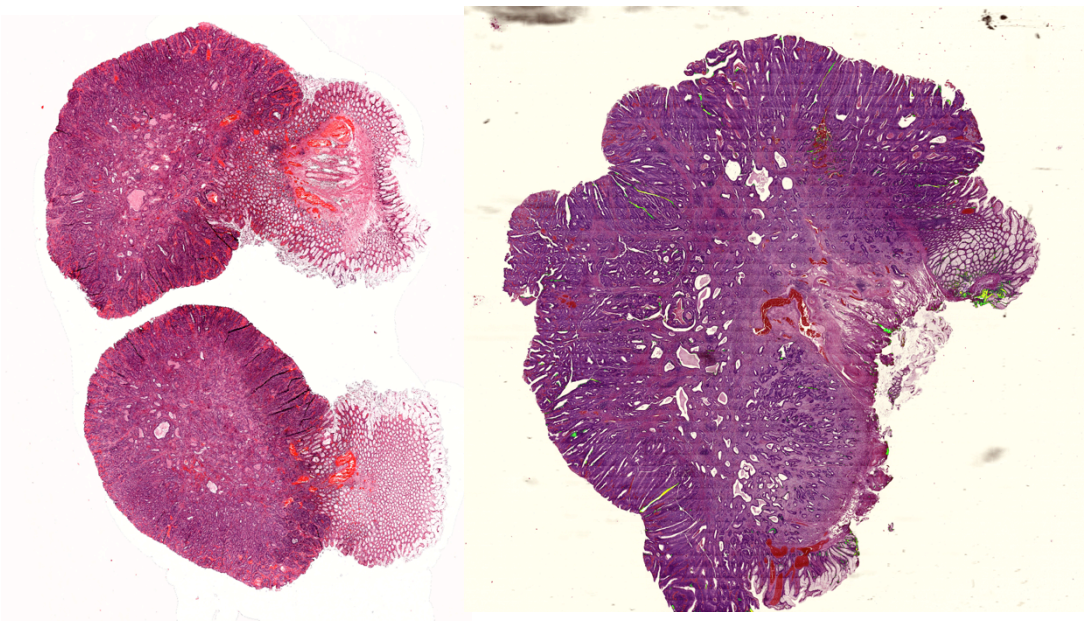


**Figure 4.4.** An example output of phase correlation stitching. This image was comprised of 26 smaller segments and took half an hour to stitch together [124].

The first automated imaging system utilised by this research was an Aperio T2 (San Diego, USA) scanner and was operated by the Leeds Institute of Molecular Medicine (LIMM). The second device was part of the Perkin Elmer Spotlight 400 Spectrometer (Massachusetts, USA).

These systems have several advantages, such as they are invariant to mathematical assumptions and that there is almost no computational complexity associated with them. However, images acquired from the Perkin Elmer Spotlight 400 Spectrometer required a further element of processing than those taken by the Aperio T2 scanners. This was because of the aforementioned problem of using a point source to illuminate small regions (Figure 4.6 A-C).



A B

**Figure 4.5.** Examples of images acquired by automated imaging systems. A) This is an image as produced by the Leeds Institute of Molecular Medicines Aperio T2 scanners. B) This is an example of the white light images produced by the Perkin Elmer Spotlight 400 Spectrometer.

The problem of vignetting can be seen within the image tiles that were acquired by the Perkin Elmer Spotlight 400 Spectrometer (Figure 4.6 A). However, this can be corrected for and this was done by selecting an individual tile from the whole image and then using this as a reference [25]. The tile chosen as the reference was representative of a piece of the slide that had no tissue on it. The correction methodology worked on the same principles as before and it can be seen from Figure 4.5 B & Figure 4.6 that correcting in this way reduced the vignetting phenomenon. Although, this method of correction was not perfect, it does allow for data exploration techniques, such as Principal Component Analysis (PCA), to be used effectively in the further analysis of images.

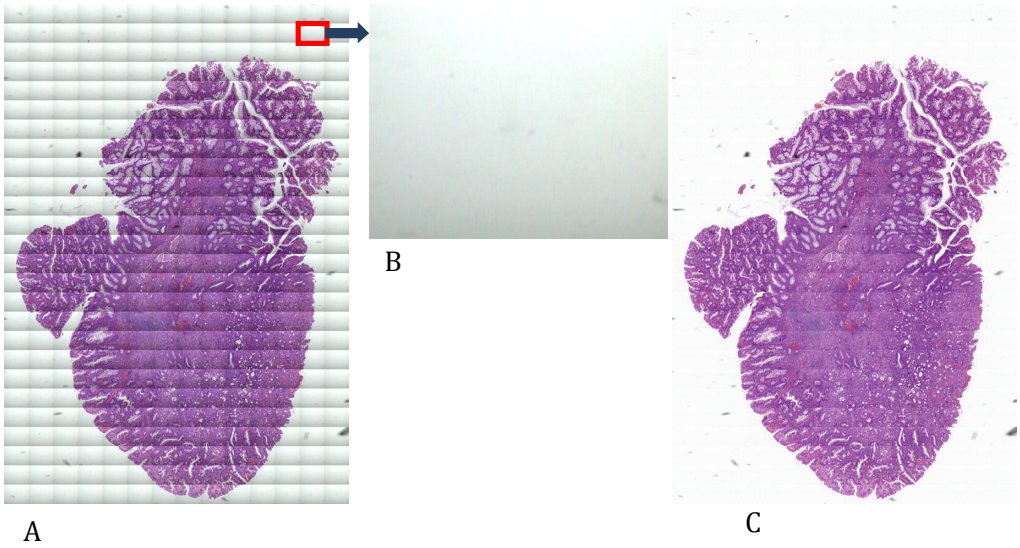



**Figure 4.6.** Correcting the vignetting phenomenon. A) The original image that was acquired by the Charge Coupled Device (CCD) of the Perkin Elmer Spotlight 400 spectrometer. B) One of the individual tiles that was used as a reference image (part of Figure 4.6 A). C) The effects of using this reference image to correct for any vignetting defects.

4.1.1.1 Linear image registration

The purpose of constructing images representative of entire tissue sections was so that they could be used further within the construction of 3D volumes. However, the images produced by stitching and tiling cannot be used directly. This is because when tissue samples were processed there was a loss of orientation between the consecutively cut sections and because some of them were deformed [24], [25] [29].

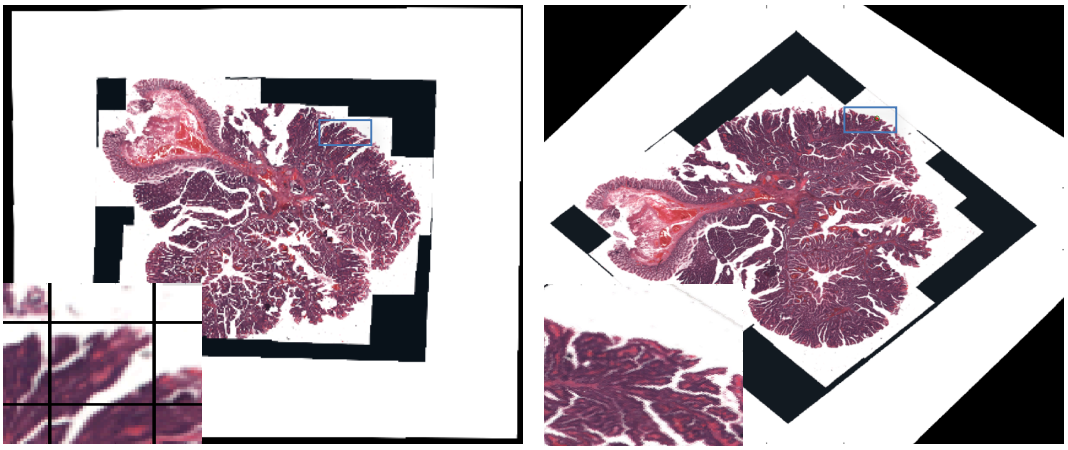
Therefore, registration algorithms were implemented. Image registration refers to the alignment of images with respect to one another so that similar features overlap [29], [32], [36], [38], [40], [44], [45], [56], [147–149]. There are many ways of achieving this, such as by linear and non-linear transformations. Using linear registrations in the attempted alignment of images involves applying rigid body transformations, which are made up of rotations, translations and scaling factors. In this instance, linear registration was facilitated by the computation of SIFT descriptors and by using the block matching algorithm.

The block-matching algorithm operates by initially splitting one of the images, referred to as the moving image, into tiled regions and this is compared to another that it is to be registered to, known as the static. The individual tiles are moved within the local neighbourhood of the static so that points of correspondence can be identified. The points of correspondence defined are the centroids of the moving images tiles that maximised the NCC when compared against those of the static image [24], [34]. It

should be noted that, normalised correlation is not the only similarity method that can be used and a review of other techniques is provided within the following reference [34].

The rigid transformation estimated by the block matching algorithm was the result of optimising the minimum Euclidean distance between points of correspondence [134]. However, it was decided that this technique was not fit for purpose. This is because the method was practically cumbersome and could not handle large degrees of rotation between images (Figure 4.7). In Figure 4.7, it can be seen that the matched regions between images are not the same.

The rotational dependence of this algorithm is related to how the images are tiled. If there is a large degree of rotation between images then the features contained within the tiles from one image will not be the same as the other. Therefore, the similarity metric will not be able to accurately define where the points of correspondence are. Although, the way normalised correlation is calculated means that correspondence points will be generated.



A B

**Figure 4.7.** The dependence of the block-matching algorithm on images being correctly rotated. A) An image that was to be registered to (B). B) The image that A is to be registered too. The zoomed locations are the areas that the block matching algorithm determined as being the same.

To reduce the impact of rotation it has been suggested in the literature that a multi iterative approach be taken [24]. In this way, it is thought that as the images become better aligned the fidelity of the correspondence points will increase. However, this will introduce a large computational burden and will mean that the volume rendering process will take place over many weeks. Therefore, this research sought rotationally invariant methods and implemented the SIFT algorithm as an alternative.

Rotational invariance in SIFT is achieved by virtue of its calculation. The algorithm finds key points within images through the construction of a Difference of Gaussian pyramid (DoG). Maximum and minimum points within this pyramid were found and their location was used to collect pixel intensity and orientation information from the original images [35]. To make the algorithm rotationally invariant the orientation information that was gathered was translated with respect to the most dominant and was included within a feature vector. The matching of points between images was then facilitated by the computation of the minimum Euclidean distance between these feature vectors. The regions identified by this algorithm were very accurate, as shown in Figure 4.8, but contained many erroneous points of correspondence.

However, these can easily be removed via the RANdom SAmple Consensus (RANSAC) algorithm [37]. RANSAC operated by iteratively fitting a model, which in this instance was composed of translations, rotations and scaling factors to matched correspondence points. It rejected any points that did not fit the optimised model to a high degree of accuracy. The refinement by RANSAC on the SIFT points can also be seen in Figure 4.8. The green circles with blue crosses are the points of correspondence that RANSAC preserved and the green circles with no crosses are the points that were rejected. From observing Figure 4.8 it can visually be seen that the matched points are strongly related and thus justifies the use of this algorithm in refining SIFT points.

The linear model as estimated by RANSAC was used further as a starting point within a simplex search optimisation algorithm [134]. This was so greater refinement between the points of correspondence could take place (Figure 4.9). The optimisation metric used within the simplex algorithm was the L1 estimator, and is presented in equation 12 and 13.

However, linear transformations are not appropriate for histology images. This is because of the non linear deformations that occur to tissue samples when they are being sectioned [25], [29], [147]. Therefore, non-rigid transformations have to be applied to images to correct for this. The non linear changes between images are often not obvious to the naked eye but could affect further analysis, as the continuity of structures will be disrupted in three dimensions.

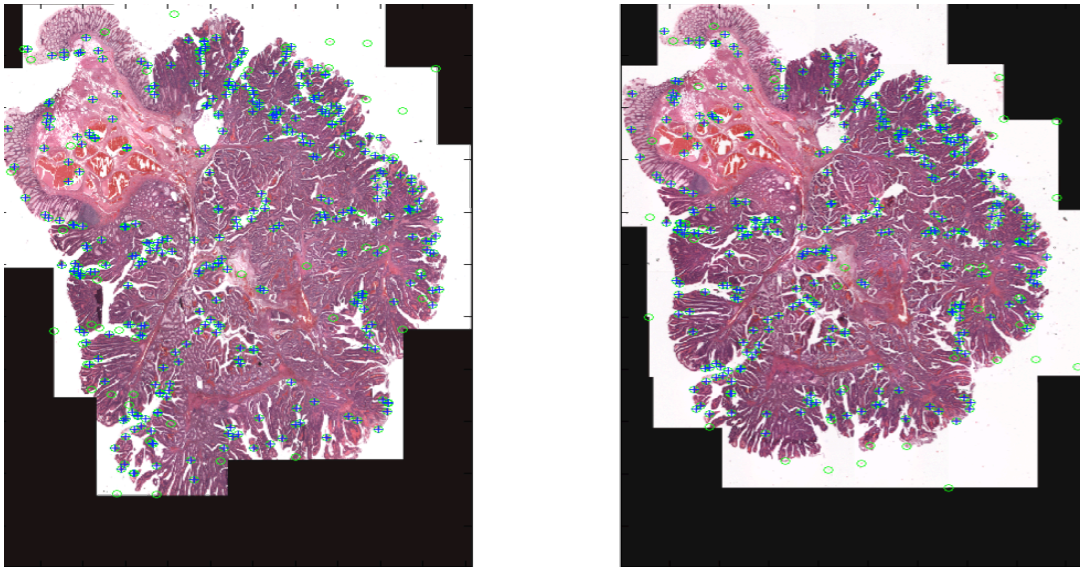
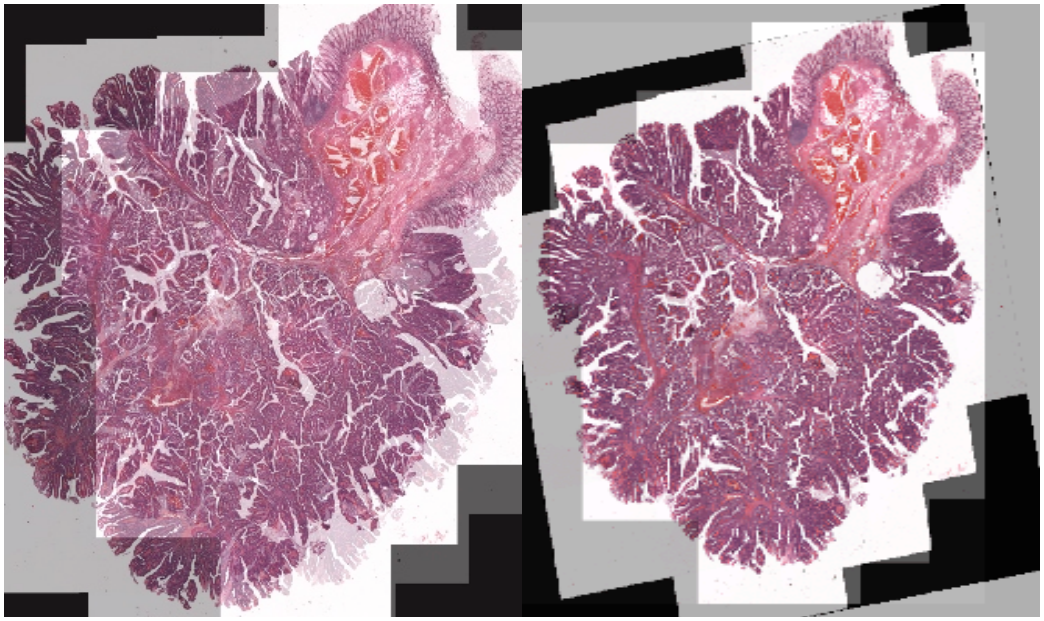


**Figure 4.8.** The SIFT points that have been identified between two images that are to be registered. The green circles with blue crosses are those points that the RANdom SAmple Consensus (RANSAC) algorithm preserved. Those without blue crosses are the points of correspondence that were rejected.



A B

**Figure 4.9.** The improvement in overlap which can be achieved by using the SIFT and RANSAC algorithm when optimising a linear transformation. A) The overlap of the raw images. B) The overlap of the linearly registered images.

4.1.1.2 Non Linear registration

As the linear transformations were not appropriate, non linear algorithms were considered. However, in some instances, the preliminary computation of a linear

transformation was necessary as it reduces registration error. Although, certain techniques that use closed solutions with correspondence points do not usually require these [31].

There are many non rigid techniques to choose from [40] but only as rigid as possible techniques are appropriate for histology images [29], [31], [40], [42], [45]. This is because they prevent the over warping of images and maintain their true characteristics. In this instance, two as rigid as possible registration schemes were investigated. These were the Moving Least Squares (MLS) algorithm and the B Spline grid deformation technique [29], [31], [42].

The MLS technique required points of correspondence as an input and in this instance RANSAC filtered SIFT points were used [29], [32], [36]. To prevent RANSAC from removing points of correspondence that were correctly matched but did not fit its estimated model well, because of deformations, the SIFT points were clustered into 4 regions. RANSAC was then performed on each cluster separately. This follows on from a similar idea within the literature whereby RANSAC was performed on matched points within individual scanning electron microscopy tiles that were part of a larger image [36].

The remaining correspondence points were then used with the rigid implementation of the MLS algorithm in order to warp one image so that it looked like another. MLS deforms images by firstly fitting a grid to them and then moving the vertices of this grid so that the points of correspondence between images optimally overlap. For the sake of accuracy, the number of vertices should be high but implementing this on a standard desktop made the algorithm computationally inefficient. This research found that a good compromise was to use a 15 by 15 grid in the deformation of images.

However, it was determined by experimentation on a large data set of images that the success of this technique in accurately aligning images was highly variable. It is evident from Figure 4.10 A - D that this procedure can be used to bring about the accurate overlap of images but Figure 4.11 A-D reveals its limitations. The unreliable nature of this registration can be attributed to the quality of the images used and the way in which erroneous correspondence points were removed.

The size, and therefore quality, of the images had to be reduced before they were analysed. This reduction was necessary because of computational constraints that apply when working with images in Matlab (Natwick, USA) on a standard desktop PC. The image resizing affected the number and quality of correspondence points found between images that were to be registered and ultimately made the task of registration more difficult.

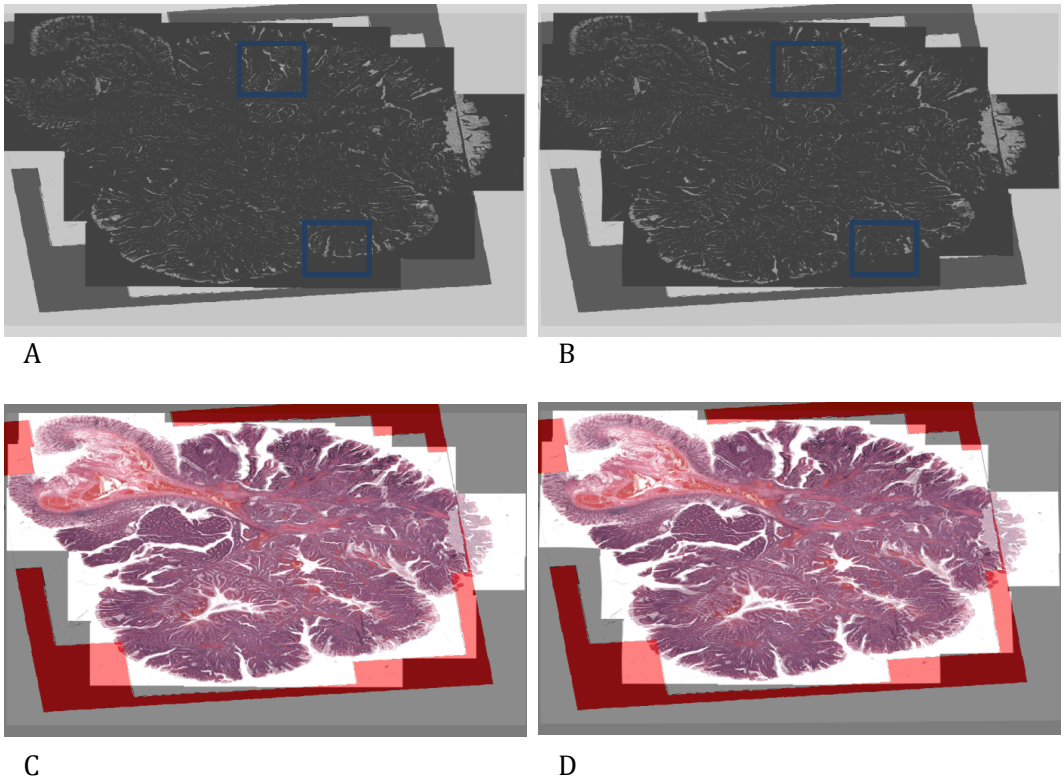



**Figure 4.10.** The improvements in alignment that can be achieved by using grouped RANSAC filtered SIFT points with the Moving Least Squares algorithm (MLS). A) The difference in overlap between images which have been linearly registered through the optimisation of SIFT points. B) A difference image made from images that have been non-linearly registered via the MLS algorithm. The blue boxes in A & B highlight regions within the images where the difference has been reduced by application of the MLS algorithm. C) A representation of the overlap of the linearly registered images. D) The overlap of the images registered by the "as rigid as possible" MLS technique.

It may have also been the case that the grouping together of SIFT points into 4 clustered regions was not appropriate. This is because it increased the likelihood of retaining erroneous correspondence points when the linear transformations of RANSAC were applied to these smaller groups.

Further problems associated with this technique relate to the assumptions used during its operation. In some instances, only a few sparse SIFT points were found for particular regions and this is because the images were resized before they were processed. It was thought that the points found in the resized images would represent the deformations that occurred to all features. However, the complex nature of histological processing meant that this was unlikely to be the case.

However, these techniques do have some potential. The further implementation of this methodology within similar research will require large-scale images to be acquired and retained. To efficiently process these images a way of performing the MLS algorithm on multiple cores for multiple image regions has to be developed.

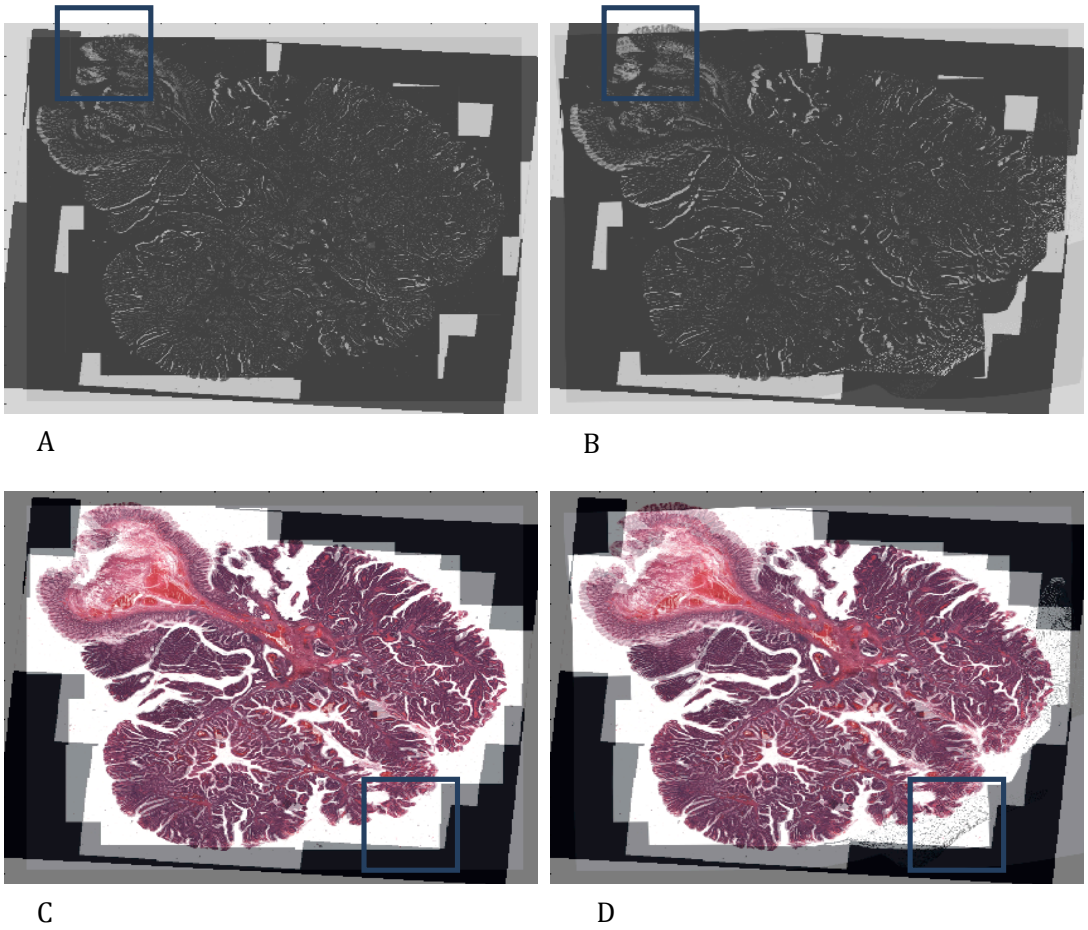



**Figure 4.11.** The error that can be introduced into images by the MLS algorithm. A) The difference image that have been registered by linear means. B) The differences between images that have been warped using the MLS algorithm. The blue boxes in A & B highlight regions where the difference has been made worse by the MLS technique. C) The overlap between images that have been linearly registered. D) The overlap between images that have been deformed using the MLS algorithm. The blue boxes in C&D highlight regions where the overlap has been made worse by applying the MLS algorithm on group RANSAC corrected SIFT points.

The technical difficulties associated with accurate in house registration led to a collaborative project with the Leeds Institute of Molecular Medicine (LIMM). Their registration algorithm has been proven within the current literature and they have the added ability of being able to access powerful automated image scanners [41–43].

Their registration algorithm works on the same principles as those presented above. It operates by initially rigidly transforming acquired images of whole tissue sections. These images are then split into tiles and these are used with their adjacent counterparts in the definition of correspondence points that were used with the non linear image registration algorithm.

The points of correspondence found are a result of maximising the phase correlation coefficient [44] and these are used with a B Spline grid to enable one image to look like another. Example registration results for their algorithm and the accuracy that can be achieved are displayed below in Figure 4.12.

In total the LIMM volume rendered 23 cases for this research project. These volumes were used further to enable the difference in continuity between epithelial structures within adenomas and cancers to be understood.

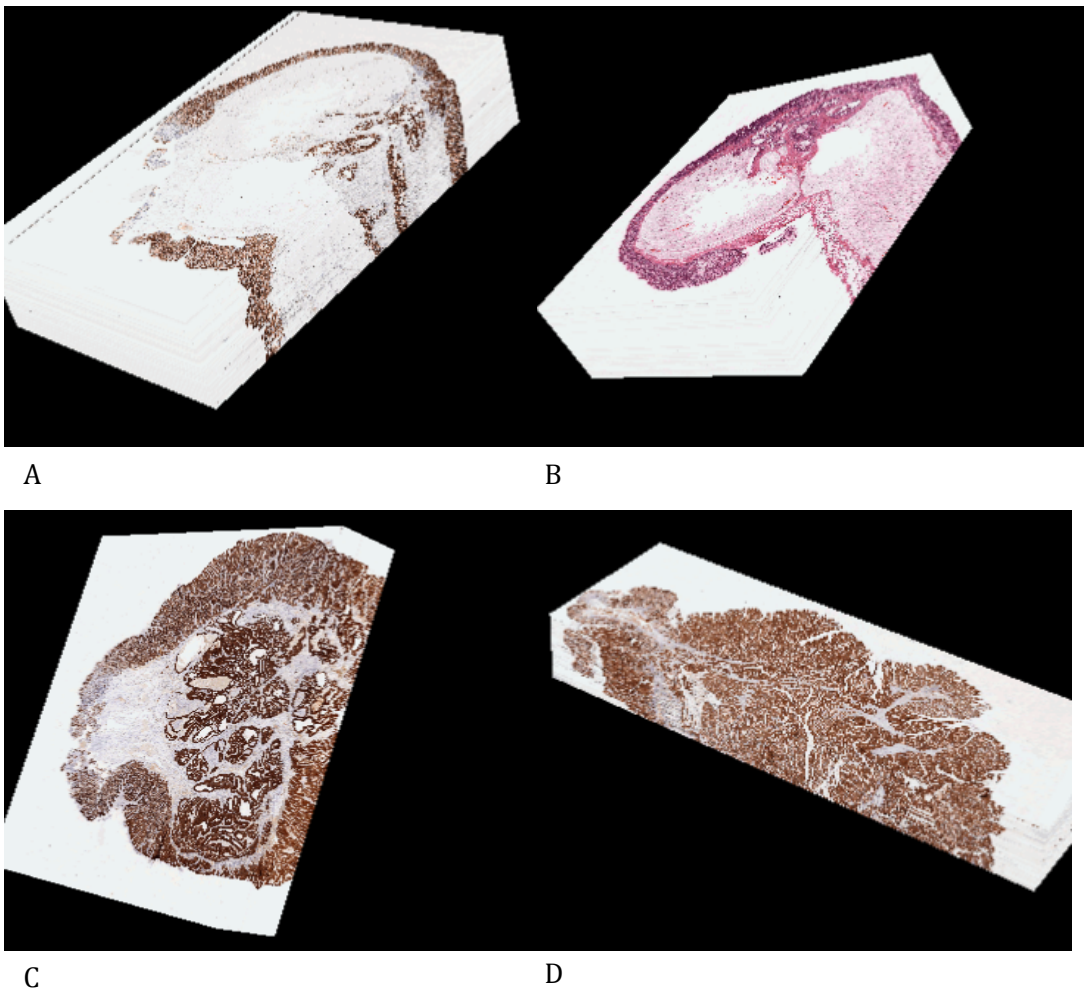



**Figure 4.12 A-D**. Example Haematoxylin and Eosin, H&E, (B) and Cytokeratin (A,C-D) volumes created for this project by the Leeds Institute of Molecular Medicine (LIMM).

### 4.1.2 Image Segmentation

The 3D volumes presented in Figure 4.12 can be used in a variety of different ways, including the subjective measurement of continuity by a pathologist. However, to quantify any differences further processing of these volumes was required. In this instance, image segmentation algorithms were investigated and developed so that the continuity of epithelial glands in 3D could be measured. As with the registration procedures, there are a plethora of image segmentation algorithms available to choose from within the current literature. The majority of these techniques can be broadly split into two types; those that use intensity information of single pixels within an image and those that do not.

#### 4.1.2.1 Pixel and texture based image classification

Intensity based techniques are important in terms of histology because staining is used in order for the structural morphology of a tissue sample to be highlighted. Therefore, by analysing differences in colour between differently stained regions of images, a feature of interest can be segmented.

Within histology the Haematoxylin and Eosin (H&E) staining reagents are most commonly used in the identification of features. They cause negative and cytoplasmic compounds within tissue samples to stain blue and red respectively. As the epithelium contains a high density of nuclei it will have a strong staining affinity for haematoxylin. The surrounding connective tissue, the submucosa, will stain strongly for eosin as it contains many cytoplasmic compartments. Therefore, the difference in colour between the epithelium and its surroundings will mean that it should be easily segmented.

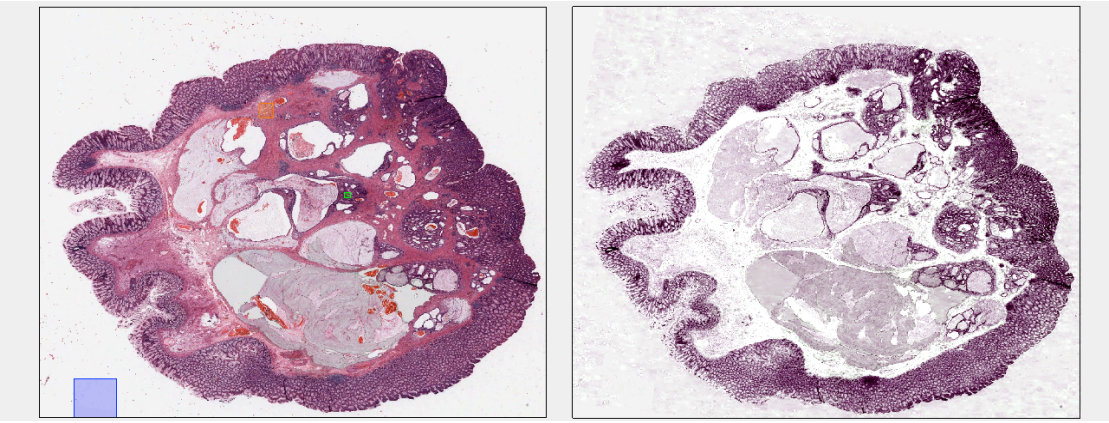
A popular stain separation algorithm within the literature is colour deconvolution and it was utilised in this research for the purposes of segmenting the epithelium from within 3D volumes [59], [60]. Colour deconvolution requires the definition of desired and undesired colours within an image along with the colour of its background.

These vectors were used to reshape the images Red, Blue and Green (RBG) basis vectors into new colour spaces, from which the undesired colour was removed. Further to this, the pixels were projected back onto the original RBG colour scheme so that only the desired colour remained (Figure 4.13).

However, as is evident from the above figure, H&E images could not be used to accurately segment the epithelium from within images and this is because these reagents are not specific for it. Other non-epithelial tissue structures will also stain the

same colour with H&E and so will be preserved along with the epithelium when colour deconvolution is applied.

Although, it can also be seen from the results of colour deconvolution, that the non-specific H&E staining was sparse relative to the epithelium. Therefore, it was thought that spatially corrected Fuzzy C Means (FCM) clustering would correct for this staining defect. FCM is a texture segmentation technique which classifies pixels based upon their and their neighbours distance away from the mean of a cluster [65]. In this way, it was hoped that the incorrect staining within the submucosa would be correctly classified.



A B

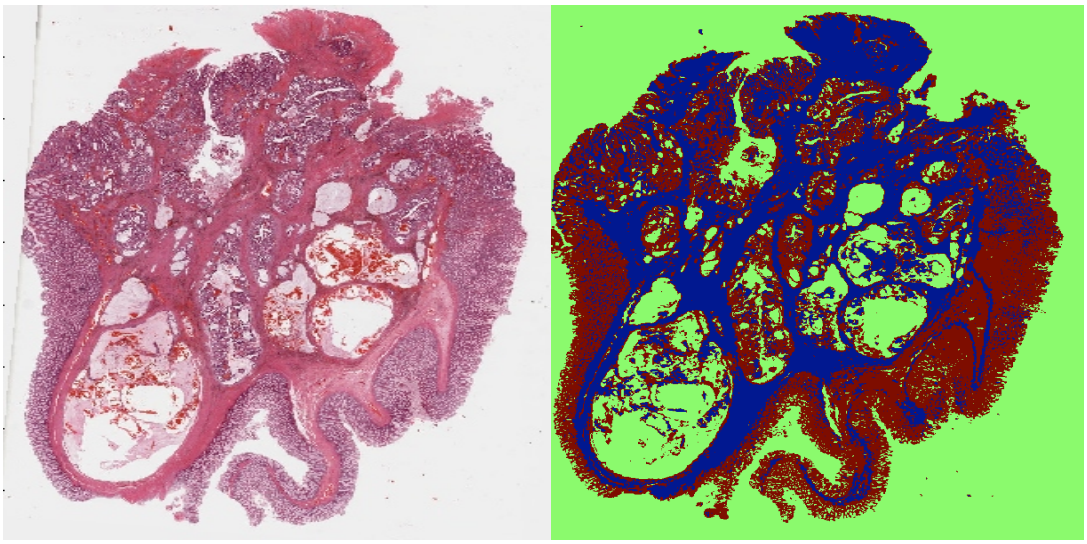
**Figure 4.13.** The colour deconvolution algorithm. A) The original H&E stained image with the desired colour contained within the green box, the undesired colour in the orange box and the background colour is contained in the blue box. B) The result of reshaping the RBG basis vectors into a new space and then deleting the undesired staining component.

However, as is evident from Figure 4.14, this was not the case. This is because more than a few pixels in the erroneously stained submucosal regions have the same colour as the epithelium. Therefore, there was an increased likelihood that they will be classified as having epithelial like staining characteristics when FCM was applied.

Problems with clustering techniques in general are that the algorithms are supervised and require the number of clusters to be defined before it is implemented. Therefore, if the incorrect number of clusters is chosen then the features of interest will not be preserved within the final classification results.

In the current literature, FCM has been more commonly used as a starting point on which other classification algorithms can improve upon[64], [65], [150]. An example

within the literature is the use of FCM with active contours without edges. However, on the contrary to this, this research found that FCM did not provide a good enough initial classification and therefore led to the evolution of erroneous contours.



A B

**Figure 4.14.** The attempted classification for the epithelium using spatially corrected Fuzzy C Means (FCM) clustering. A) The original H&E image and B) the FCM classified image.

To investigate the potential of using active contours, the colour deconvolution algorithm was implemented as an alternative to FCM. The desired and undesired colour components that were required for colour deconvolution were determined from performing PCA on entire image stacks. This allowed for similar staining patterns to be identified throughout a volume as PCA had been simultaneously performed on all of its images. In this instance, the data matrix used with PCA contained the voxels within its rows and their intensities within its columns, Figure 3.3. However, PCA will not allow the intensity of staining colours to be automatically derived, it will only allow for their identification.

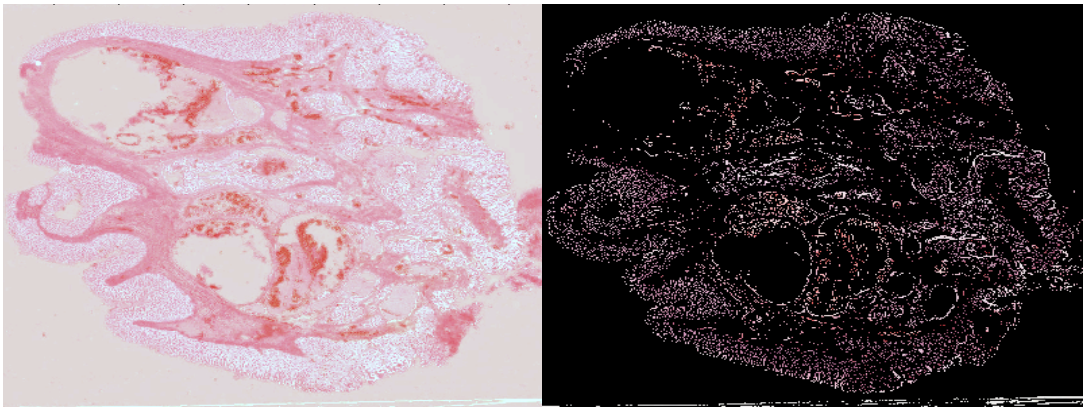
Therefore, appropriate thresholds were used with a test score image so that the staining colours the components were identifying could be quantified. These thresholds were used further with the rest of a volume's images so that colour deconvolution could be enacted throughout.

The sign of the thresholds used was determined by the direction of the PCA score in which the feature of interest was represented. If a feature of interest had a high score

value then pixels were selected that were below the maximum of this score and vice versa for those that had low values.

When using a procedure of this sort, multiple images will be generated that will relate to the different deconvolved stains. In this investigation, it was determined that deconvolving the eosin staining colour provided the best initialisation for active contours. This was because it left behind white holes in epithelial positions, Figure 4.15 A, which when used with a Gaussian filter facilitated the production of seed points, Figure 4.15 B.

Gaussian filtering caused the white holes in images to have lower intensity values. As a consequence, epithelial regions could be identified by subtracting the filtered image from the original. Within the subtracted images epithelial regions had negative values and so a simple threshold could be used to select initialisation points for use with the active contours with out edges algorithm.



A B

**Figure 4.15.** How deconvolving the eosin stain can be used to locate epithelial regions. A) The eosin deconvolved image. The eosin staining colour was selected from PCA components that had been performed on an image stack. B) The result of subtracting the Gaussian filtered deconvolved image away from the original. These regions were then used as seed points for active contours.

In this instance, active contours were grown using the principles of contraction and dilation. This technique uses a series of rules to decide whether the growth of a contour into a particular region is appropriate, please see section 3.1.3.3. The technique is extremely efficient at finding object boundaries, as can be seen in Figure 4.16, and it was thought that its use with epithelial seed points would enable this feature to be segmented.

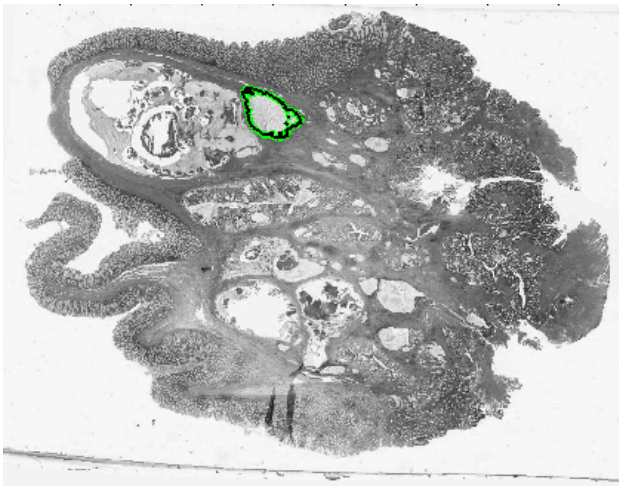


**Figure 4.16.** The accurate segmentation of objects that can be achieved using active contours without edges that are evolved by contraction and dilation.

However, as can be seen from Figure 4.17, the retention of erroneous seed points within Figure 4.15 B was problematic. They caused contours to grow in non-epithelial regions and meant that this feature could not be segmented in this way.

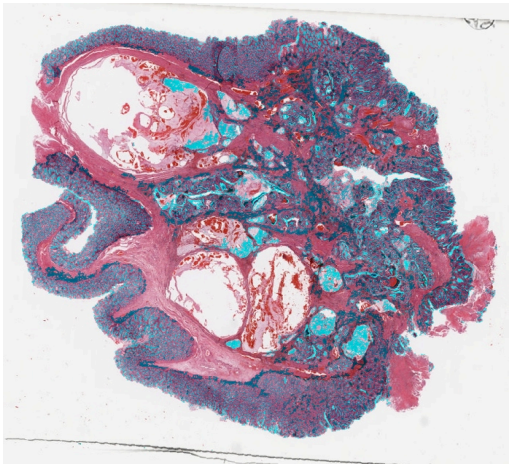


**Figure 4.17.** The end result of using eosin colour deconvolution initiated active contours.

Attempts at removing these undesired contours began with using the colour of the pixels they enveloped in their classification. However, this was practically cumbersome to implement and did not work in all instances. Therefore, alternative segmentation methods were sought. However, it should be noted that other active contour procedures exist and that only one was investigated here [65] ,[63], [151].

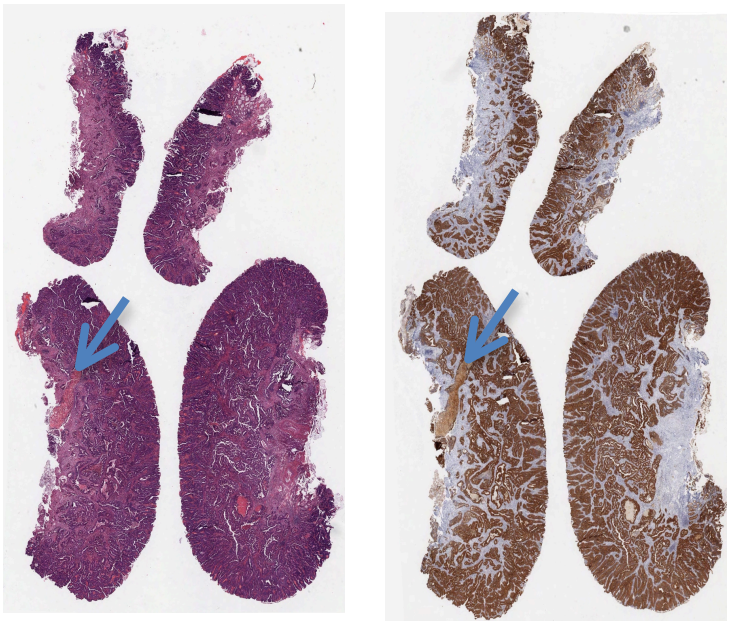
4.1.2.2 Classification by intermodal registration

Because of the problems associated with abstracting the epithelium from within H&E images, immunohistochemistry (IHC) alternatives were considered. In this instance,

the pan Cytokeratin (CK) antibody, MNF116 (Neomarkers, UK), was used as this antibody is very specific for the epithelium (Figure 4.18 B). However, the staining of the tissue sections epithelial regions by the CK antibody is still prone to experimental error. These errors include the incorrect dilutions of antibodies by technicians and non specific antibody binding. An example of this erroneous staining is highlighted by the blue arrow in Figure 4.18 B.

The blue arrow in Figure 4.18 B indicates where a blood vessel within the CK section has stained a similar colour to the epithelium. This is problematic, as the use of the CK staining colour in segmentation will mean that these features will also be preserved. However, by combining the two images, the blood vessel staining in the CK image can be corrected for and the epithelium can be segmented.

This is because in the H&E image, Figure 4.18 A, the blood vessels have stained red and the epithelium has stained purple. Therefore, the registration of the CK and H&E images will mean that histology specific staining patterns can be identified. For example, purple and brown colours, due to the H&E and CK images respectively, will represent epithelial pixels.



A B

**Figure 4.18.** The non-specific staining of a tissue sections blood vessels, indicated by a blue arrow on B, by the MNF116 pan Cytokeratin antibody.

The exploitation of these staining relationships allowed for accurate epithelial segmentation. This involved serially sectioning tissue blocks and alternatively staining them with H&E and CK.

Initially, the different sets of stained images were registered separately using the algorithm developed by the Leeds Institute of Molecular Medicine (LIMM) [41], [42]. This enabled for the similar image sets to be registered with respect to one another. Corresponding images from each data set were then selected and used within the Consensus Principal Component Analysis (CPCA) algorithm [137]. CPCA looks for common trends between measurements that have been made for a system, via the use of weighting coefficients. In this instance, the epithelial staining relationship identified above was represented within the CPCA scores and this facilitated its segmentation.

This technique required that for every CK image there was a corresponding H&E image. In some instances, the histological processing meant that individual CK or H&E tissue sections were destroyed [27]. Therefore, there were no images of alternatively stained tissue sections for these cases. So that correspondences could be found an automated algorithm was developed that enabled image pairs to be identified, Figure 4.19.

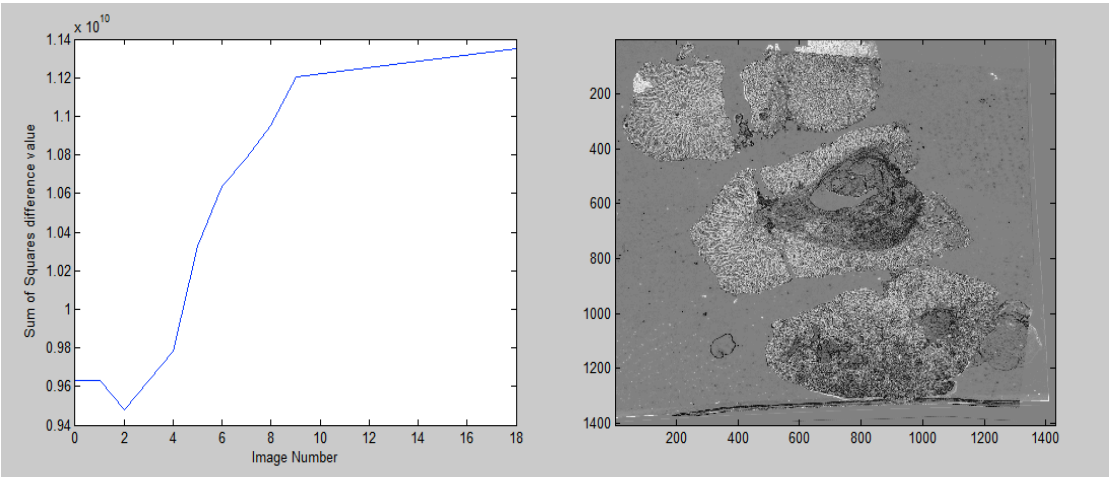



**Figure 4.19.** The results obtained from the automated image selection algorithm. A) This graph displays the Sum of Squares Differences (SSD) [28] between H&E images that were registered with respect to a target CK image. B) A depiction of the overlap between the corresponding sobel images found by this algorithm.

For this automated pairing, an image was selected from the CK volume for a sample and, where possible, the nine images before and after its corresponding position

within the H&E volume were deformed with respect to it. Deformation in this instance was carried out between sobel edge filtered images [28] as this enabled for similar regions between images to be made more obvious.

Initially, non rigid deformations during image matching were brought about by using subsampling and by using a 2 by 2 B Spline grid. Once the Sum of Squares Differences (SSD) [28] between images had been optimised at a particular resolution and B Spline grid size, the subsampling was reduced and the number of splines was doubled. This continued until both the maximum resolution of the images and the algorithms maximum number of splines, which was 8 by 8 in this instance, had been reached. Registration stopped when a threshold was passed or when the algorithm exceeded a set number of iterations [29].

Once corresponding images had been found, a more refined registration took place. In this instance, a 16 by 16 B Spline grid with a low convergence value was used within the refined registration. The correspondence algorithm was implemented for all images and enabled for accurate correspondences to be found in over 2,800 images. However, in some cases the predicted correspondence was poor and on these occasions images were deleted.

From these registered images, regions representative of a tissue sample's epithelium, blood vessels, smooth vascular tissue and slide background were selected and included within a training set image, Figure 4.20.

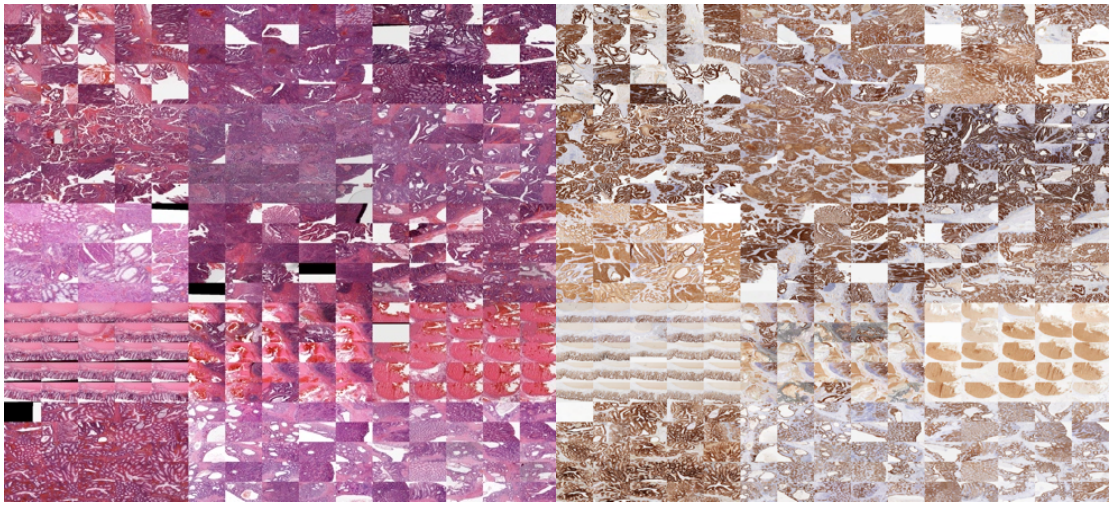


**Figure 4.20.** The training set of registered images that were used in the identification of the epithelium from within intermodally registered volumes.

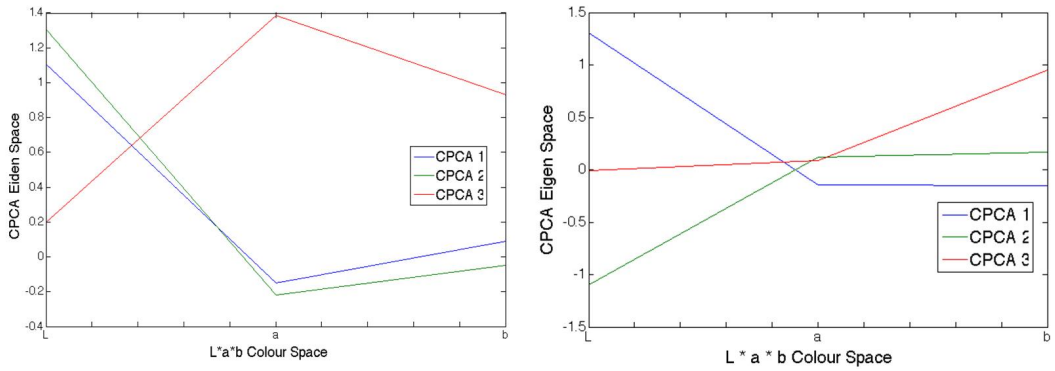
CPCA was conducted on these training images and the algorithms first three loadings and weights were used to derive epithelial specific scores from each samples

intermodally registered volume. In this instance, the automated image pairing algorithm discussed previously was used to combine a sample's H&E and CK images together. This resulted in each sample having a volume that was composed of alternatively stained sequentially ordered images of tissue sections.

Since all images of a samples volume had a complementary CK or H&E counterpart, the CPCA loadings derived from the training dataset could now be applied to it. This facilitated the production of CPCA superscores where the epithelial regions contained within the co-registered images were highlighted. For each volume, a dataset of epithelial specific superscores were chosen and used with the 90% confidence limits of the D Statistic so that a binary epithelial specific volume could be made.

The CPCA loadings derived from the training set of images are presented in Figure 4.21 A&B. Within Figure 4.21 A it can be seen that the first CK image block loading is describing the difference between light and dark pixels. For the second component dark staining that has high blue values was being described. This is in keeping with the blue staining of the submucosa within images of CK stained tissue sections. The third component relates to brown staining and this was because the L*a*b luminance value is low and because the blue axis is in the yellow direction.

For the H&E image, Figure 4.21 B, it can be seen that, in the first component, light colours reflective of the image background were being explained. A similar relationship is associated with the second H&E image block loading. The similar features identified between these might be representative of the best match CPCA could make between the H&E and CK image variances. Within the third component it can be seen that dark reds are being described and this was representative of an image’s blood vessels and its associated components.




A B

**Figure 4.21.** The CPCA block loadings for the training set CK image (A) and the H&E image (B). The CPCA weights for the three CK image block loadings were 0.7576, 0.6493 and 0.5178 respectively. For the H&E image the three CPCA block loadings weights were 0.6527, 0.7605 and 0.8555.

Further analysis of the training images CPCA superscores in a 3D plot, Figure 4.22, gives greater confidence that the above identified trends are being used in the segmentation of the epithelium. For the first CPCA component it can be seen that the transition from white to dark pixels was being described. This component allowed for the localisation of the epithelium and this is evident from the fact that there was a strong association between brown and purple staining for this feature within this component.

The second CPCA component described the overlap between the blue submucosal pixels in the CK image with the purple/pink staining colour for this feature in the H&E image. For the third CPCA component it can be seen that the red and brown staining for the blood vessels within the respective H&E and CK images was being described.

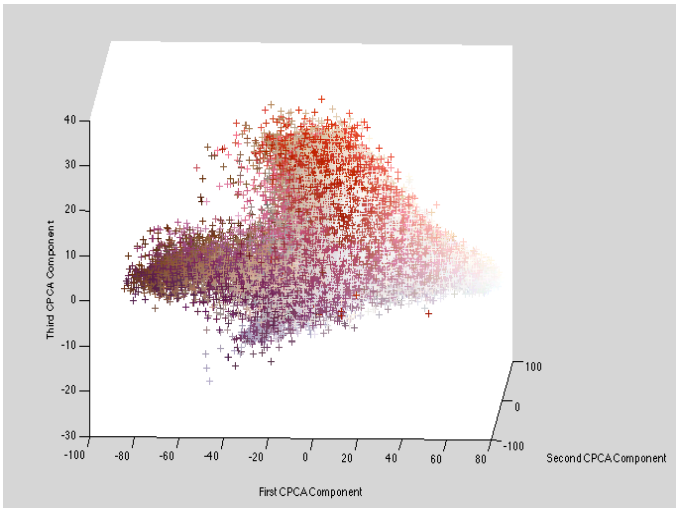
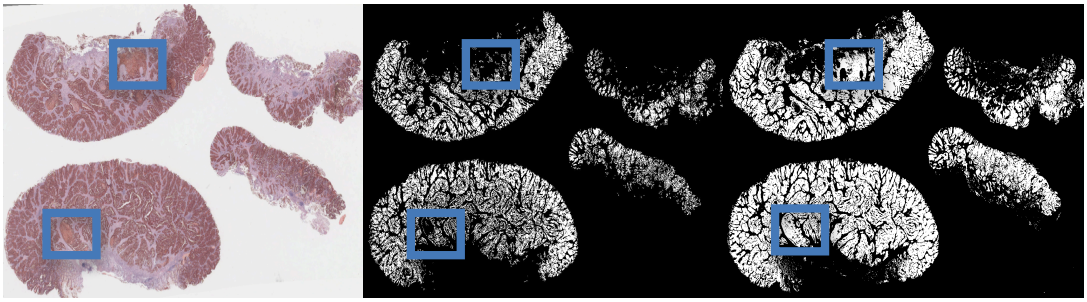



**Figure 4.22.** The 3D CPCA super score plot made from the training set of images presented in Figure 4.20.

The benefits of analysing the H&E and CK images by CPCA can be seen in Figure 4.23. In this image the overlap of the H&E and CK images is presented along with two segmentation results. Figure 4.23 B represents when the fused images were used in the segmentation and it can be seen that blood vessels, highlighted by the blue boxes, have not been preserved. Figure 4.23 C was made from using PCA with the CK image in the segmentation process and it can be seen that the blood vessels in the blue boxes have been preserved.

The combination of CK and H&E images also has other benefits such as removing anomalies present within the background slide, where for instance the edge of the glass slide was sometimes present (Figure 4.24). However, this approach had its limitations and required that for both the CK and H&E image there be strong staining.



A B C

**Figure 4.23.** Improvements in segmentation by using fused images. A) The combined H&E and CK image for a sample of cancer. B) The segmentation results for when the intermodal staining relationships were modelled by CPCA. C) The segmentation for when only the CK image was used with PCA.

In one case of cancer, the H&E staining was poor and the acquired images of this stain contained almost no features, Figure 4.25. This meant that the segmentation algorithm developed in this research could not be used in the accurate localisation of the epithelium. Therefore, no further analysis was conducted on this case.

However, the segmentation did work for the vast majority of samples and this technique was preferable to the others investigated. Therefore, the use of CPCA in the retention of epithelial features was justified.

Within Appendix A of this thesis the first segmented image from within each analysed volume is presented. Within Electronic Appendix A, .gif files are presented which demonstrate the accuracy of this method in preserving the epithelium from within intermodally registered volumes. In Appendix F, the pathological classification for each movie is presented.

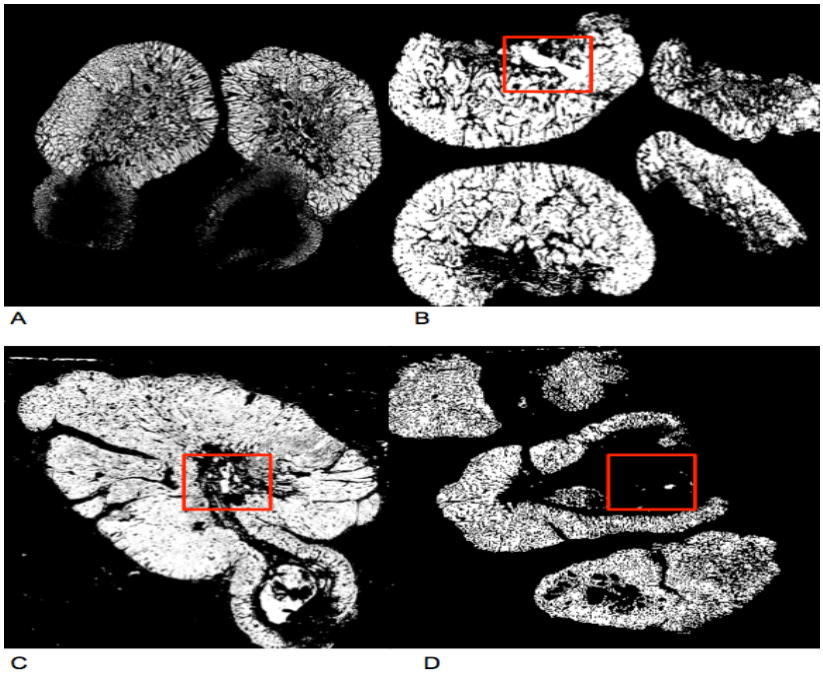



**Figure 4.24.** The segmentation results when only the CK images were used in the segmentation process. A) An example of an image data set that was well segmented when only the CK images were used. B-D) Poor examples of when only the CK images were used in the segmentation process. Those areas in red are erroneous features that have been retained even after the 90% confidence limits of a training set of data had been applied

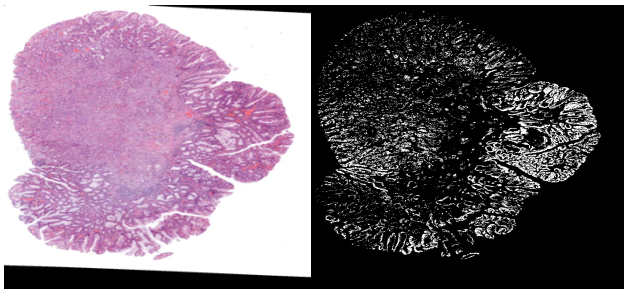
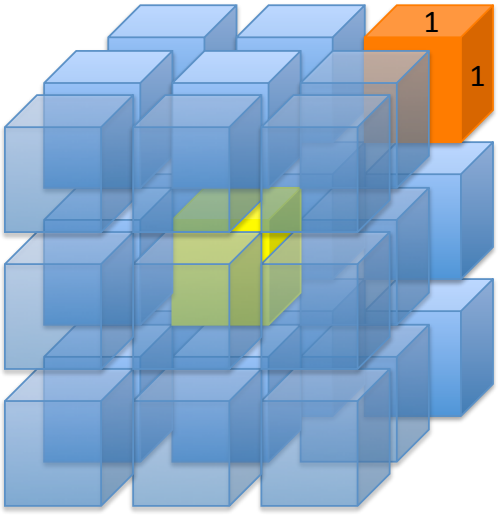


**Figure 4.25.** The poor H&E staining for a case of cancer meant that the segmentation for this sample was less than optimal.

4.1.3 The continuity of epithelial structures in 3D

For this study, the number of 26 connected components associated with the epithelium specific volumes was recorded as this is a robust measure of continuity. A 26 connected component example is provided within Figure 4.26.




26 neighbours surrounding a non zero voxel. In this instance, the middle voxel, coloured in yellow, would be sad to be connected to the positively valued voxel coloured in orange.

**Figure 4.26.** A diagrammatic representation of the continuity used by this research to investigate if there were any significant differences between how the cancer and Epithelial Misplacement (EM) cases were connected.

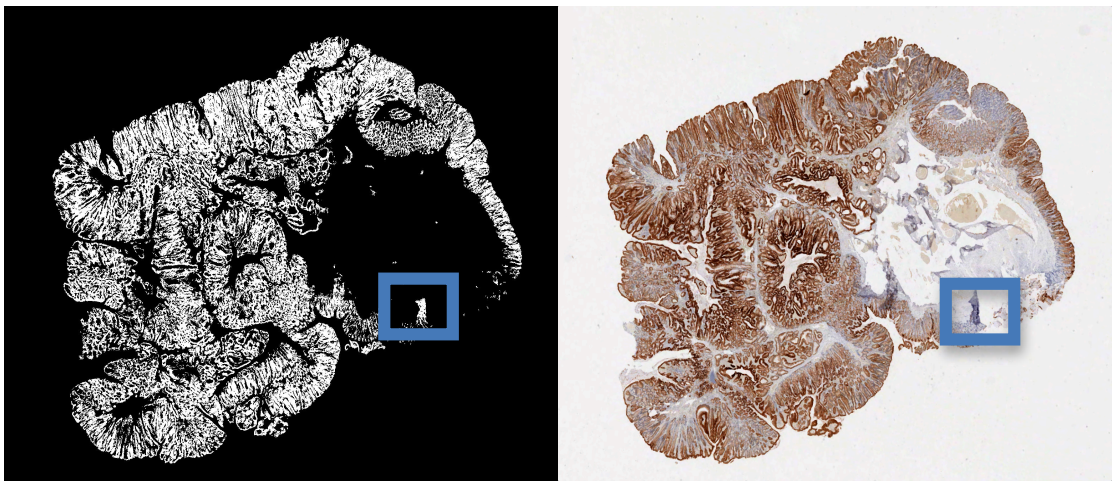
In the figure above, the middle voxel, coloured in yellow, will be said to be connected to the orange voxel because it falls within the yellow cubes 26 voxel neighbourhood. Within the epithelial specific volumes, if there are more independent voxels within the cancerous volumes, i.e. the volume contains a large number of connected components, this will mean that their submucosal glandular tissue is more independent of the surface. Therefore, this can be used as a measure of connectivity and will enable the benefit of 3D volume rendering to bring about the differentiation of EM and cancer to be evaluated. The number of connected components recorded from within these volumes is presented within Table 4.1.

To determine any differences using the defined metric, the Mann Whitney U test was used. This tested the null hypothesis that there was no difference in connectivity between cases of EM and cancer. From using the numbers in Table 4.1 no significant differences could be found and therefore the null hypothesis could not be rejected.

**Table 4.1.** The number of 3D connected components found from the segmented volumes presented in Electronic Appendix A.

| Pathology | Connectivity Index | Pathology | Connectivity Index |
|---|---|---|---|
| **EM** | 5129 | Cancer | 2269 |
| **EM** | 2406 | Cancer | 611 |
| **EM** | 1458 | Cancer | 997 |
| **EM** | 1487 | Cancer | 3367 |
| **EM** | 642 | Cancer | 3133 |
| **EM** | 2724 | Cancer | 1304 |
| **EM** | 2557 | Cancer | 1563 |
| **EM** | 3002 | Cancer | 2666 |
| **EM** | 3386 | Cancer | 3740 |
| **EM** | 1493 | | |
| **EM** | 1345 | | |
| **EM** | 3614 | | |
| **EM** | 1689 | | |

However, this does not necessarily mean that cases of cancer are as connected as cases of EM. This is because the analysis presented above has several limitations. For instance, in some images, there was uncharacteristically dark staining for the submucosa within CK images. This caused these pixels to fall within the 90% confidence limits of the D Statistic and meant that these features were preserved within the final segmented volume, the blue box in Figure 4.27.



A B

**Figure 4.27**. An example where the epithelium has been segmented along with other image artefacts. The erroneously preserved region is contained within the blue box.

Other problems included misregistration in epithelial regions and this meant that the staining relationships identified above could not be used to accurately demarcate the feature of interest. Also, in some cases, the presence of glue on slides, used to attach

cover slips to histological sections, was sometimes preserved within the final segmented volumes. This is because it produced extremely dark colours within the acquired images.

However, these features are essentially random events and so voxels from the 3D segmented volumes that were only one level thick were removed. This is a very coarse quality assurance technique and so it is not surprising that it did not work well in all instances.

A review of the training set superscores did not identify any components relating to these anomalies and so they could not be removed using the training set of images presented in Figure 4.20. However, the inclusion of specific examples of these within the training set is unlikely to remove all undesired features, as they will not be representative of all. Also, the identification of these undesired features prior to segmentation is difficult and so the results presented here should be treated as a high level segmentation upon which other algorithms can improve upon.

Given all of these difficulties it is obvious that this segmentation technique will have to be further refined before any definite rejection of the null hypothesis can take place. Although, it should be noted that the results presented are a vast improvement upon those that can be found within the current literature. In these algorithms, glandular tissue is often missed or other histological components are misclassified [67], [71], [116], [118]. In this analysis, all of the epithelium can be segmented but with the addition of undesired features in some cases.

### 4.1.4 Summary

The above analysis suggests that the submucosal epithelial islands in the cases of cancer are as connected as their EM counterparts. This is because there were no significant differences between the number of 26 connected components recorded for epithelial specific volumes, as presented in Table 4.1.

It may have been the case that this investigation's experimental design was floored but the segmentation technique developed has several advantages over the other methods tested, Figure 4.23 and Electronic Appendix A. Therefore, the use of this approach was justified.

This technique was highly specific for the epithelium and does not preserve large undesired image artifacts, Figure 4.24. It is also preferable to others that are presented within the literature because in these cases the presented results demonstrate that histological components are not always accurately identified [65], [67], [71].

It is thought that the cancer cases used in this research might be representative of very early stage cancers and therefore they might have not yet become independent of the surface. Although, the 3D volume rendering approach was found not to be helpful, several techniques were developed that had application with other modalities. For example, image stitching allowed images of whole tissue sections to be made [123] and these could be deformed [29] with respect to Infrared (IR) images, section 4.3.

The next chapter describes the IHC alternatives investigated by this research. This was used further with the Mann Whitney U statistic to determine if any significant differences existed between the staining intensity of EM and cancer cases.

### 4.2 Immunohistochemistry (IHC)

In the following chapter an indication of the diagnostic potential of E-Cadherin, p53, Collagen IV, Ki67 and MMP-1 will be given. In this instance, the scores given to the Immunohistochemistry (IHC) tissue sections by a Senior Histopathologist were evaluated using a two tailed Mann Whitney U test. This only provided information on the discriminatory potential of the individual antibodies and so hidden trends contained within the data could have been easily missed. Therefore, unsupervised Hierarchical Cluster Analysis (HCA) was implemented in an attempt to identify these discreet trends.

Within the literature, antibody annealing has been used to successfully differentiate Epithelial Misplacement (EM) from cancer [9], [11–13]. In these studies, the unique genetic code of proteins have been utilised so that their expression within different regions of a sample can be visualised via antibody annealing.

The intensity of the annealing reaction gives information on how strongly expressed a particular protein is. For example, if there is weak staining in a region then this means that there is a reduced expression of this protein at the location of interest. The differences in staining intensity between samples are what researchers have used previously in the determination of the characteristics that belong to EM [9], [11–13].

In this study, ten EM and cancerous cases were investigated using the antibodies identified above. This was so the claims of the literature could be scrutinized [11–13]. Although, in reality, a large cohort of sample will have to be used before any solid conclusions can be drawn.

In this investigation, a double review of the staining intensity of the investigated samples was conducted. An expert's assessment, who has experience in determining tissue samples that are representative of EM, was used along with those of this author in the determination of differences in staining between samples. In all cases agreement between the two grading's was noted.

In this instance, assessment took the form of comparing the sample's submucosal glandular tissue components against their surface epithelium. Any differences between these two histological components were noted and the results of the analysis are presented within this chapter.

In this investigation, differences between the E-Cadherin, p53, MMP1 and Ki67 antibodies were quantified solely by intensity of staining. For Collagen IV, differences

in the continuity of staining for this antibody were noted between the basement membranes of the surface and submucosal glandular tissue.

It should be noted that for the rest of this chapter, commentary on differences in staining are with respect to the submucosal epithelial tissue. For example, if it is noted that staining has decreased then it refers to the fact that relative to the surface mucosa there was less intense staining in the submucosal epithelial tissue.

### 4.2.1 Comparisons with the literature

In previous research, it has been determined that in the EM cases that there was reduced MMP-1 and p53 staining within submucosal glandular tissue[13]. For cancer, it has been found that there was reduced submucosal epithelial staining for the E-Cadherin and Collagen IV proteins [13]. These findings will be compared against this investigation and each antibody is presented within its own section below.

#### 4.2.1.1 E-Cadherin

The epithelial staining of E-Cadherin in cases of EM is presented within Figure 4.28. In this figure, six samples are displayed and it can be seen that there is strong uniform staining for both the submucosal and surface glandular tissue. In these cases, the submucosal epithelial tissue is located below the blue lines presented on Figure 4.28. These are typical examples of EM and for all cases it was noted that the intensity of staining for the submucosal epithelial tissue was the same as the surface.

For the cancer cases, there was increased staining for E-Cadherin within the submucosal glandular tissue within 50% of samples. However, it should be noted that differences in staining between the submucosal and mucosal epithelial tissue was minimal in most cases, if not completely the same. This is evident from Figure 4.29 A-B where it can be seen that the differences in staining between these histological features is indistinguishable. In Figure 4.29 A, the cancerous epithelium is above the blue line and the adenomatous surface tissue is below it. Within this image it can be seen that they have stained the same colour.

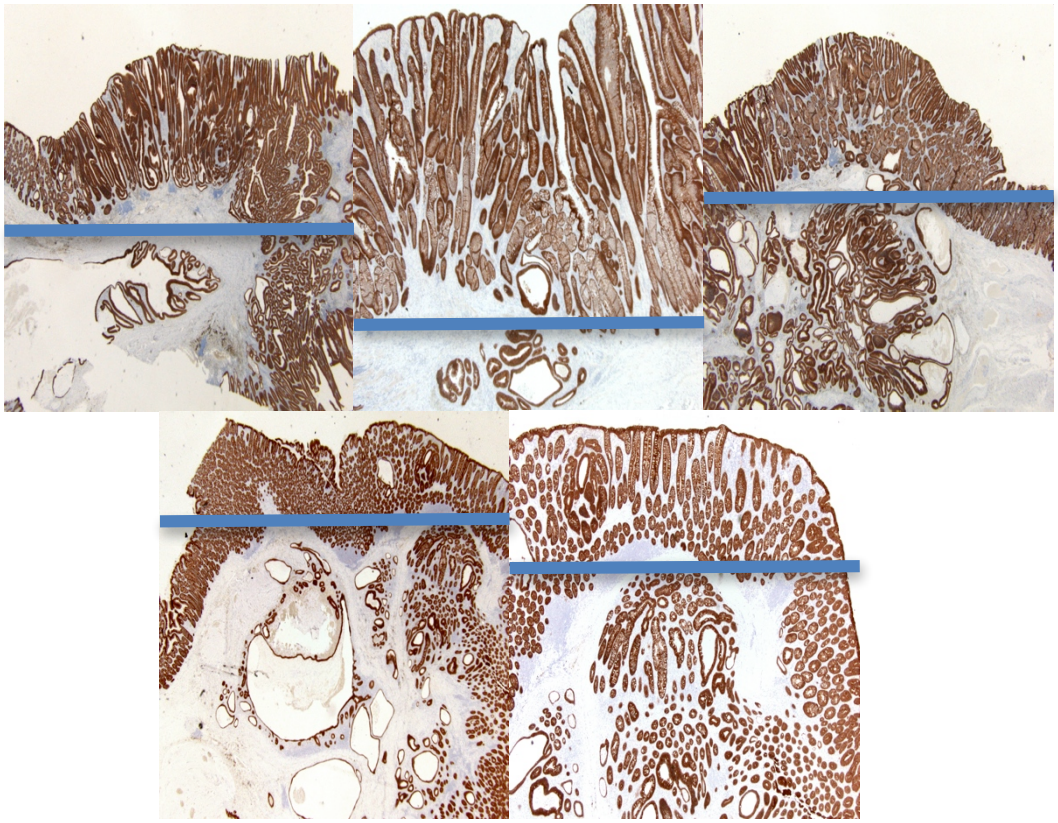
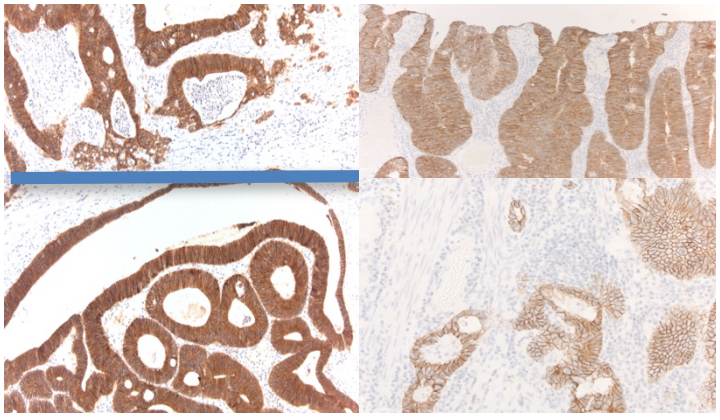


**Figure 4.28.** The E-Cadherin staining characteristics of cases of EM. It can be noted that the submucosal epithelial tissue, which are those features under the blue bar, have stained the same intensity as the surface, the tissue that is above the blue bar.

In Figure 4.29 B, the top half of this image represents the surface epithelial tissue of a case of cancer. The bottom image represents the cancerous tissue and again it can be noted that both have stained a similar intensity.



A B

**Figure 4.29.** The E-Cadherin staining characteristics for two cases of cancer. A&B) In both these images the staining characteristics for the submucosal epithelial tissue is the same as the surface.

The finding that there was increased E-Cadherin staining within the cancerous pathologies was surprising because in these cases it should be expected to be reduced. This is because this protein is related to cellular adhesion and a reduction in intensity would therefore represent less expression of this protein. This would enable the epithelial tissue to be more fluid [18]. However, it should be noted that only 65% of cases in a recent study showed decreased staining for this antibody within cancerous tissue [13].

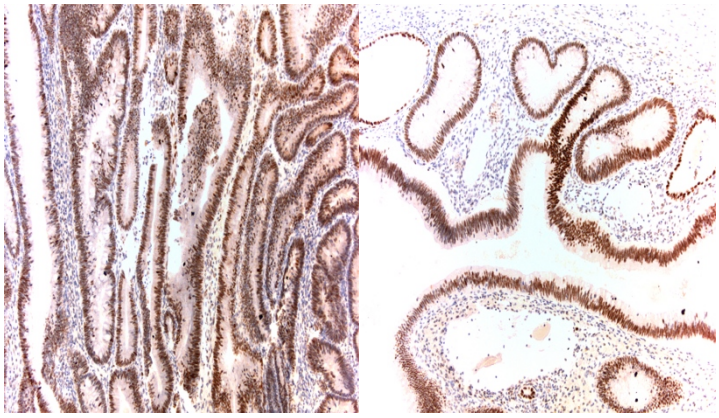
In this instance, the increase in staining intensity in the cancerous tissue was found to be statistically significant when using the 99% confidence limits of the Mann Whitney U Statistic. However, the results presented here should be treated with caution and this is because only a small sample size was used. This compiled with the fact that in over half of the cases analysed there was similar staining between the surface and submucosal glandular tissue in the cancerous cases means that this antibody is not a good marker.

#### 4.2.1.2 p53

For p53, the staining intensity for this antibody varied between strong and moderate within the EM cases. For all samples, it was found that there were no differences in staining intensity between the samples surfaces and their submucosal epithelial tissues.

This is in keeping with the current literature. It has been determined mutant variations of this protein are found within cancerous pathologies and that any alterations that occur to the p53 protein will cause its half life to increase [17]. As a result of this, this protein becomes accumulated within the nuclei of damaged cells. Therefore, an increase presence of this protein should be expected within the cancerous, but not adenomatous, pathologies.

Since the p53 antibody is a nuclear stain, it is best-visualised using high magnifications and the staining for an EM sample is presented within Figure 4.30. In this figure it can be seen that for the presented sample that there is no difference in staining between the surface epithelium, A, and the misplaced epithelium, B.



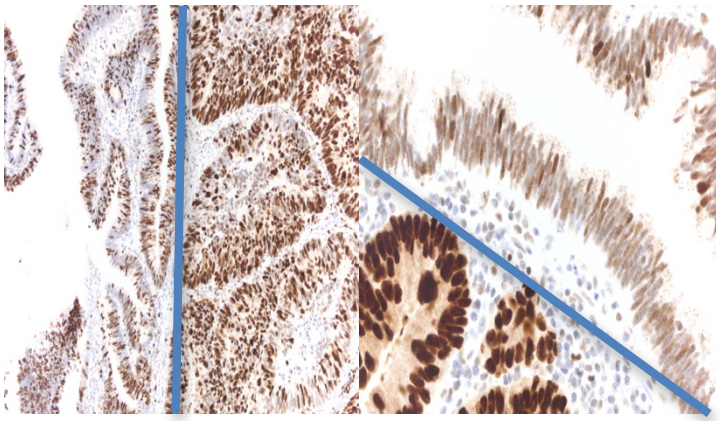
A B

**Figure 4.30.** p53 antibody staining for a case of EM. A) p53 staining within the surface epithelium. B) p53 staining within the misplaced epithelium. These images have been acquired using a x40 magnification.

For the cancer cases, there was increased submucosal glandular tissue staining for this antibody within 40% of cases. An example of this staining is presented within Figure 4.31 A&B.

In Figure 4.31 A, it can be noted that the cancerous epithelial tissue has stained with a stronger intensity than its adenomatous counterpart. A similar staining relationship can be seen in B where the cancer tissue, which is below the blue line, is a darker colour than the surface adenomatous tissue, which is above the blue line.

This is in keeping with the current literature where it was noted that there was increased staining for the submucosal glandular tissue in 61% of cancer cases [13].



A B

**Figure 4.31.** A representation of the nuclear staining for p53 within cases of cancer. A) The cancerous tissue, to the right of the blue line, can be seen to have a stronger staining intensity than the surface adenomatous epithelium, which is to the left of the blue line. B) The cancerous tissue, below the blue line, has stronger staining intensities than the surface adenomatous tissue, which is above the blue line.

However, it should be noted that in 20% of the cancer cases there was increased staining within the surface adenomatous tissue. An example of this is presented in Figure 4.32. For the rest of the cancer cases it was determined that there was similar staining between the surface adenoma and the submucosal epithelial tissue.

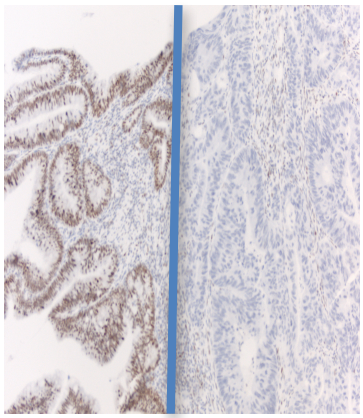


**Figure 4.32.** p53 nuclear staining for surface adenomatous tissue within a case of cancer. The adenomatous tissue, which is to the left of the blue line, shows p53 staining and the cancer tissue, which is right of this line, shows none.

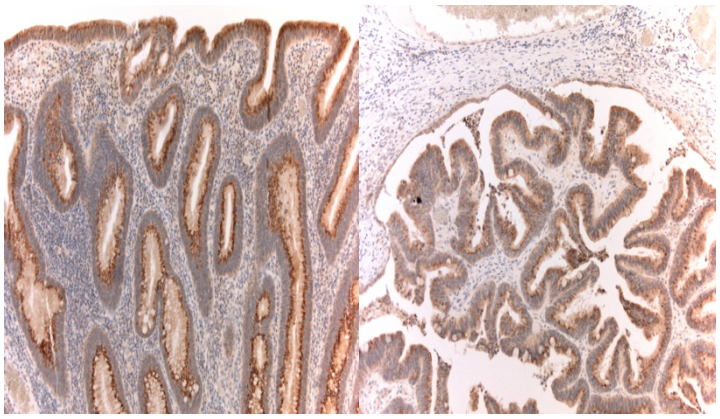
Although, there were some differences in staining in the cancerous pathologies, further analysis by the Mann Whitney U statistic found no significant differences. This is on the contrary to the analysis as published by Yantiss et al [13] but it should be noted that the results presented here should be treated with caution because of the small sample sizes involved.

4.2.1.3 Matrix Metalloproteinase 1 (MMP1)

The next protein investigated was MMP1 which is a collagenase that in cases of adenocarcinoma is thought to help facilitate metastasis[152]. In the literature, it has been determined that there is increased staining for this protein within stromal and epithelial regions [13]. In the investigation by Yantiss et al it was found that there was increased staining for MMP1 within the submucosal epithelial and stromal tissue in 39% and 91% of cancerous cases respectively [13]. However, it should also be noted that Yantiss et al found that there was also increased staining within EM cases. In their study they found that 4% of EM samples showed increased epithelial staining and 43% had increased stromal staining [13].

On the contrary to the above analysis, this investigation found that for the EM tissue samples that there was no stromal staining for MMP1. This could be indicative of the fact that this membrane degrading enzyme was not present but may also be representative of defective staining.

The lack of stromal staining is evident from Figure 4.33, which displays examples of typical EM staining. From Figure 4.33 A it can be seen that there is no stromal staining for the MMP1 antibody in the surface epithelium and Figure 4.33 B shows the lack of staining within the submucosal glandular tissue in EM samples.



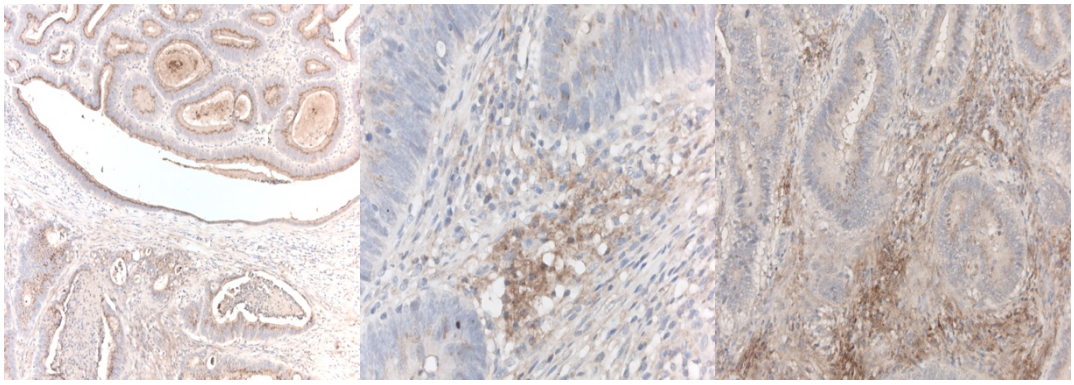
A B

**Figure 4.33.** The lack of stromal staining within EM samples. A) The lack of stromal staining within the surface adenoma and B) the lack of MMP1 staining in the submucosal glandular tissue.

For the cancerous cases, there was MMP1 staining within both the epithelial and stromal areas. Stromal regions are composed of cellular bodies that make up the connective tissue of samples and include fibroblasts and smooth muscle cells. It has been previously noted that within cancerous cases that the stromal staining for MMP 1 was associated with fibroblasts. It was thought that this was probably forming part of

the desmoplastic reaction, which is the surrounding of epithelial tissue by dense fibres [13]. In this investigation it was noted that the submucosal stromal regions had increased staining in 100% of cases.

However, it should be noted there were only minimal differences in staining intensity between the surface stromal and submucosal regions. The staining for this antibody was also highly variable. This is evident from Figure 4.34 A-C where it can be seen that there was a large degree of variable staining between samples.



A B C

**Figure 4.34** Cancerous stromal staining by the MMP1 antibody. A) The weak stromal antibody staining for cancerous cells that are contained within the bottom of the image. B) Another example of weak stromal staining. C) An example of moderate staining for the MMP1 antibody in the stroma of cancerous tissue.

In this instance, analysis by the Mann Whitney U statistics 99% confidence limits found that the null hypothesis could be rejected. However, given that there was no stromal staining in the EM cases this result should be treated with caution. This is because previous research has found staining for MMP1 within EM samples [13]. Therefore, differences seen on this occasion maybe due to poor staining.

Although there was no stromal staining in the EM samples, it can be noted from Figure 4.33 that there was some reaction for this antibody within epithelial regions. Therefore, the staining may not have been defective and the lack of MMP 1 staining for EM cases might be genuine but further investigations are required.

Analysis of the epithelial tissue in the EM samples revealed that 40% displayed decreased submucosal epithelial staining. For the rest of the samples, the submucosal epithelial tissue stained the same intensity as their surfaces.

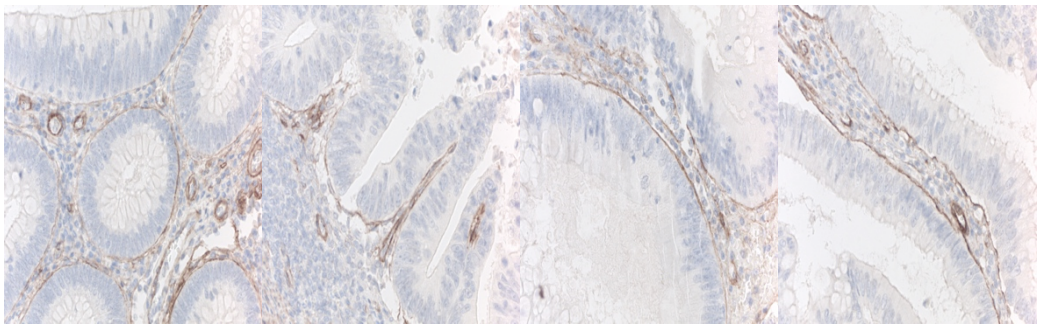
For the cancer cases, it was found that there was increased staining for the MMP1 antibody within the submucosal glandular tissue in 30% of samples. In one case it was

found that there was increased surface epithelial staining and for the rest it was the same. Analysis by the Mann Whitney U statistic found no significant differences between the pathologies. Therefore, the diagnostic use of this antibody is questionable.

4.2.1.4 Collagen IV

Collagen IV is another membranous protein that is thought to have application in the differentiation of EM samples from cancer. It is a constituent component of basement membranes and a breakdown in this membranous component is believed to be correlated with metastasis[16]. In previous research, it was noted that in the EM samples there was strong continuous staining for this antibody [13].

This fits well with this investigation's conclusions and some examples of Collagen IV antibody staining is presented within Figure 4.35 A-D. In Figure 4.35 A&B, the surface epithelial tissue is shown and it can be seen that the Collagen IV staining is continuous in both images. In Figure 4.35 C&D, the submucosal epithelial tissue is shown and it can also be seen in these images that there is continuous staining for the Collagen IV protein.



A B C D

**Figure 4.35.** Examples of Collagen IV staining within the EM samples. A&B) The continuous representation of the basement membrane within the surface epithelium for two cases of EM. C&D) Continuous staining for the Collagen IV antibody within submucosal glandular tissue.

Analysis of the cancerous cases submucosal epithelial tissue found that the continuity of Collagen IV staining was disrupted if not completely absent in 100% of cases. An example of the lack of Collagen IV staining within the cancer cases is presented in Figure 4.36. This finding correlates with what has been found previously in the current literature [13] and this antibody may have some diagnostic potential.

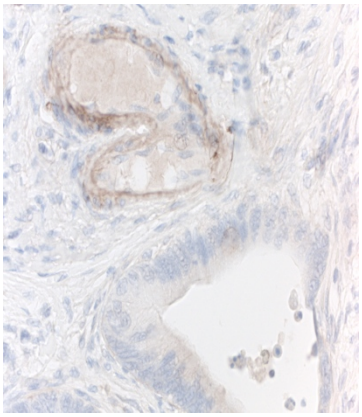


**Figure 4.36.** The lack of Collagen IV staining within the basement membrane of a cancerous epithelial cell that was found within the submucosa. For comparative purposes a blood vessel is also presented in the top of the image and it can be seen that in this histological feature that there was continual and strong staining for Collagen IV, where as there is none in the epithelium.

Analysis by the Mann Whitney U Statistics 99% confidence limits revealed that there were significant differences between the EM and cancerous pathologies for this antibody. Unlike MMP 1, this biomarker may have some diagnostic potential because there were no assumptions made about the staining characteristics of tissue during the statistical testing. However, only a small sample size was analysed here and further investigations will need to use a larger cohort before any conclusions are drawn.

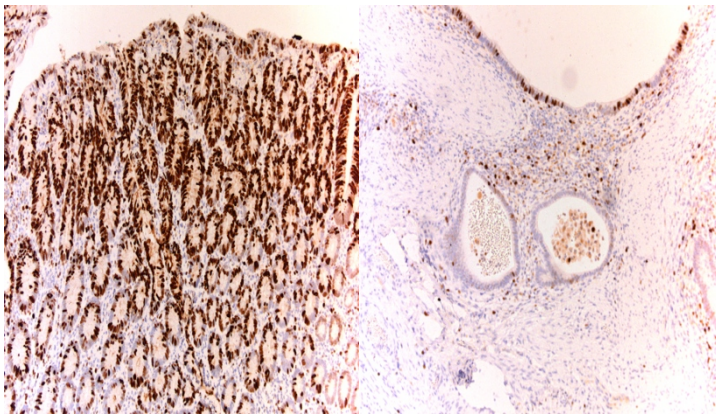
4.2.1.5 Ki67

The final antibody studied in this research was Ki67. It is a proliferation marker that has been used previously in describing the characteristics of EM [11], [12]. It should be noted that within the current literature no significance testing has been conducted for this antibody. Therefore, it is difficult to compare the results presented here against the current literature. Although, it should be noted that other research has determined that the staining of EM samples for this antibody is of little diagnostic use [11], [12].

In this investigation, there was increased staining within the cancerous submucosal tissue in 50% of samples. Within one sample there was increased staining in the surface adenoma and for the rest of the samples the staining was the same.

For the EM samples, it was noted that there was a reduction in staining intensity within the submucosal glands in 40% of cases and for the rest of the samples the staining was the same. An example of the reduction in staining that was seen in some of the samples is presented within Figure 4.37.

Analysis by the 99% confidence limits of the Mann Whitney U statistic revealed that there were significant differences in staining between pathologies and meant that the null hypothesis could be rejected. However, it should be noted that only 50% and 40% of the cancer and EM cases respectively showed any changes in staining. Therefore, this antibody may have a limited impact within a clinical setting.



A B

**Figure 4.37.** Reduced Ki67 staining within an EM sample. A) The surface epithelial tissue within a sample of EM where there is strong staining for Ki67. B) The misplaced epithelial glands where it can be seen that the Ki67 staining intensity is reduced.

4.2.2 Multivariate Analysis

The above analysis only provided univariate information on the effectiveness of the individual antibodies in discriminating EM from cancer. For some antibodies, such as Ki67, significant differences between pathology groups were seen but only a small set of samples were correctly classified. By using different combinations of the IHC staining scores it may be possible to group samples together and increase the accuracy of the individual classification.

To investigate this, Hierarchical Cluster Analysis (HCA) was conducted using a similar approach to O Stefansson et al 2011[153] . In their research, raw pathological grading scores were used within a Euclidean distance based, average linkage Hierarchical Cluster Analysis (HCA) model so that normal and carcinogenic breast tissue could be categorised into distinct groups.

For this research, it was not practical nor was it required to use all unique combinations of the investigated antibodies. This is because the scores given to some antibodies, such as Collagen IV, brought about complete classification. When using the scores for these antibodies within a HCA model, unique pathological clustering will always be found. Therefore, they will mask the discriminatory potential of the other biomarkers that were investigated.

As a result, the scores associated with particular antibodies were sequentially removed from the dataset and new HCA models were constructed. This enabled the contribution the other biomarker combinations were making towards the observed clustering to be evaluated. In the first instance, all of the immunohistochemistry scores were included within HCA. For this model, and all that follow, the matrix used with HCA contained the tissue samples within its rows and the IHC scores for these cases within its columns, Figure 4.38.

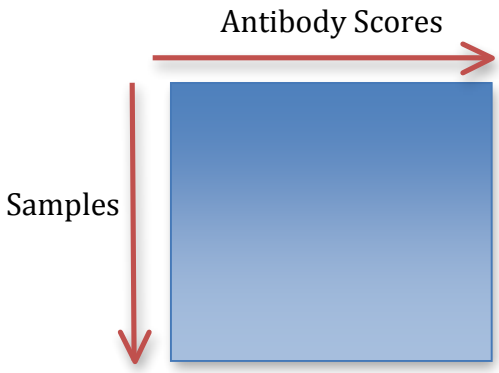



**Figure 4.38**. The format of the data entered into the HCA model.

When all antibody scores were included within HCA two very obvious pathological groupings could be found, Figure 4.39. However, since Collagen IV gave a very significant univariate result it was likely to be having a large effect on the observed clustering. Indeed, when HCA was carried out using only the scores of Collagen IV complete discrimination between EM and cancer could be brought about.

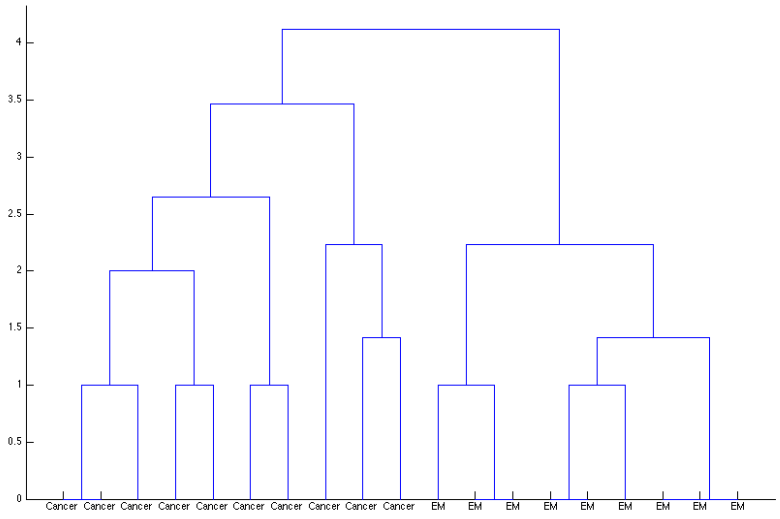



**Figure 4.39.** The clustering of pathologies by Hierarchical Cluster Analysis when all antibody scores are included within a model. It should be noted that because one of the EM samples was destroyed that this model only contains nine EM samples.

Therefore, to test the effectiveness of the other antibodies a new HCA model was made that excluded the IHC scores for Collagen IV. When excluding this antibody, it is interesting to note that the full differentiation of the pathologies did not take place. This is surprising because this research found that MMP-1 stromal staining was significantly different between the pathologies groups. Therefore, it would be expected that this would have a large impact on the HCA model and that the full differentiation of the pathologies should have take place.

Within this model two of the cancerous cases have uniquely and independently clustered within the EM samples, Figure 4.40. For these cancer cases, no differences in staining between the mucosal and submucosal epithelial tissue for E-Cadherin, p53 and Ki67 could be found. This was on the contrary to the rest of the cancerous dataset and so they would be closer to the EM samples within the IHC scoring feature space. However, it should be noted that they form their own distinct clusters within the EM samples and so a good classification is still being brought about.

Although inclusion of MMP-1 stromal staining in the model below did not uniquely cluster the pathologies, it is still likely to be playing the most significant role. This is because it enabled 100% of the cancerous samples to be differentiated from their EM counterparts. For the EM samples no stromal staining for MMP-1 was found in either the surface or submucosal glandular regions. However, within the literature it is noted as being present [13] within the submucosal stromal features of EM. Therefore, the inclusion of this staining characteristic within the HCA model below is questionable because defective staining could have skewed this researches result.

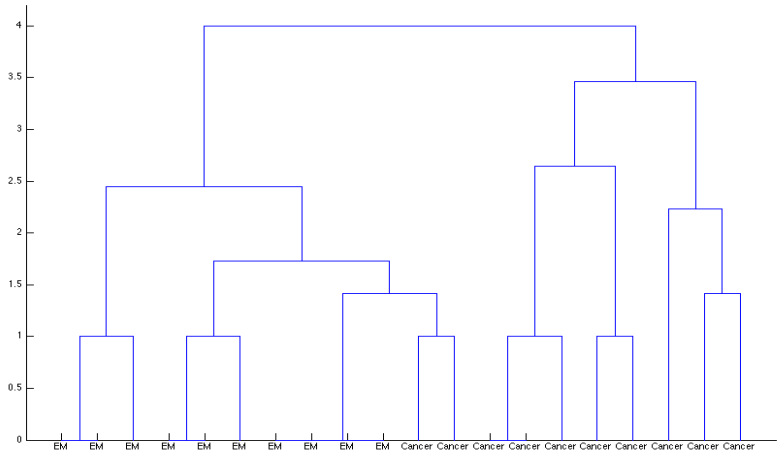



**Figure 4.40.** The results of HCA clustering when only the pathology staining scores for E-cadherin, p53, MMP-1 (both stromal and epithelium) and Ki67 were included.

As a result, MMP-1 stromal staining was removed and a new HCA model was built. The result of doing so is presented in Figure 4.41 and it can be seen from this that 60% of the cancer samples were independently clustered from the EM cases. Within the other major cluster, two adenocarcinomas are uniquely grouped from the rest of the EM samples. The unique cancer clusters are coloured red in Figure. However, given the small sample sizes involved with this study, a larger cohort of antibodies will need to be investigated before any significance can be attributed to this classification.

The results below give an indication of the underlying pathology specific immunohistochemistry staining trends that exist within the dataset. A review of the staining scores associated with the first group, which contains two EM samples and one case of cancer, would suggest that decreased submucosal epithelial MMP-1 staining was important for their characterisation. These cases also had similar surface and submucosal E-Cadherin, p53 and Ki67 staining intensities and the aforementioned characteristics were unique to this group.

The next two independent EM clusters were formed because both of these samples had decreased submucosal staining for Ki-67. These did not form a unique cluster because of differences in how their epithelial regions stained with MMP-1. Within one of the samples, the surface epithelial tissue stained more intensely with MMP-1 than the glandular tissue found within its submucosa. In the other EM case these histological features stained with the same intensity.

The next cluster, made up of four EM samples and one cancer case, is comprised of those pathologies where there were no differences in staining between their submucosal and mucosal epithelial tissue. The next two EM samples as found within Figure were characterised by the fact that they had decreased submucosal Ki67 staining. For the other antibodies included within this model, the IHC staining scores for these EM samples were the same.

The rest of the above dendrogram is made up of the cancerous pathologies. The first two cancer cases were being clustered together because these were the only adenocarcinomas were there was increased surface MMP-1 staining. The next two cancer cases clustered together because these two were the only samples that demonstrated more intense mucosal staining for E-Cadherin, p53 and Ki67. The cancerous sample adjacent to this group was very similar except in the case of p53 where no differences in staining between the samples surface and submucosal glandular tissue could be found.

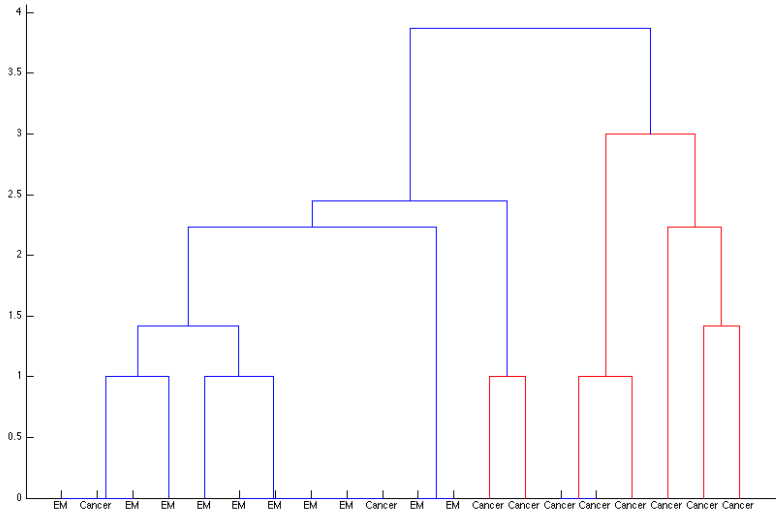



**Figure 4.41.** The HCA clustering results when Collagen IV and stromal MMP-1 staining was left out of the model. The distinct cancer groupings have been coloured red and the EM/Cancer groupings have been coloured blue.

For the remaining cancer groups, all showed decreased surface p53 staining but increased mucosal Ki67 staining. The isolated adenocarcinoma, third from the right hand side of Figure 4.41 is due to the fact that although this sample showed the former staining regime, it displayed decreased surface epithelial staining for MMP-1.

In subsequent analysis, different combinations of the IHC staining scores recorded for the epithelial staining of MMP-1, Ki67, p53 and E-Cadherin were included within new HCA models. In the first instance, leave one out antibody combinations were used with HCA. This allowed four different models to be built, due to the four choose three property of the chosen combination, but none of these improved upon the classification seen above. Therefore, no further discussion for these is provided.

In this research, it was found that p53 and MMP-1 epithelial antibody staining did not bring about significant differentiation between EM and cancer. Therefore, a HCA model was made that included only those remaining antibodies that brought about significant discrimination, which in this case was E-Cadherin and Ki67. This model is presented within Figure 4.42.

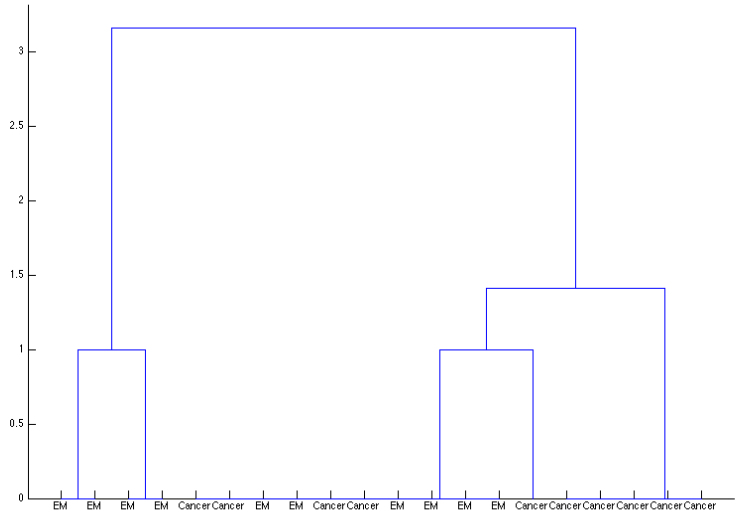



**Figure 4.42.** The HCA model made from the EM and cancer samples E-Cadherin and Ki67 staining characteristics.

From the above model, it is evident that 40% of the EM samples cluster independently of the rest. Within the other cluster, 50% of the cancer cases can be uniquely differentiated from the EM samples. The cluster that uniquely contains 40% of the EM cases has formed because these samples contained reduced submucosal staining for Ki67. The unique cancer subcluster, found on the right hand side of Figure, has formed because it relates to those samples that have increased submucosal glandular tissue staining for Ki67.

The final HCA model that will be discussed is the one that contains scores for p53 and epithelial MMP-1 staining. This model was investigated because although these biomarkers did not bring about significant univariate differences, they could potentially be used together to bring about differentiation. From Figure 4.43 it is evident that some pathology specific clustering can be achieved. However, the small sample sizes associated with this research means that this analysis should be treated with caution.

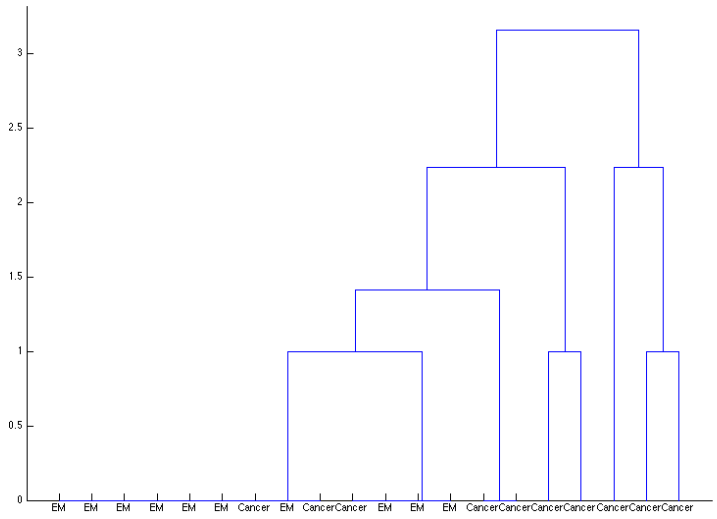



**Figure 4.43.** The EM and cancer clustering achieved when using the staining intensity scores for EM and cancer.

Within the above figure it is evident that 30% of the cancer samples can be uniquely differentiated from the rest. Although the above dendrogram contains two antibody scores, the clustering that was being observed can be explained by the univariate classifications that were discussed previously. For instance, the carcinomas found to the right of the dendrogram are those samples that had increased staining for p53 within their submucosal glandular tissue.

In all the models built subsequent to this, the only classification that was being achieved was due to single antibodies, i.e. no further increase in classification accuracy was being brought about. Therefore, no further commentary is provided.

### 4.2.3 Summary

The above investigation demonstrates that great caution should be taken when interpreting the results of the literature. In a recent publication it was found that the investigated caused significant differences in staining [13].

However, on the contrary to this, this research found that only Collagen IV could be used reliably (Figure 4.35 & Figure 4.36). This is because no significant differences in staining were found for either the E-Cadherin or the p53 antibody. Although, stromal staining by MMP 1 caused significant differences in staining, there was a complete lack of stromal staining with this antibody in the EM samples (Figure 4.33). The literature found the opposite of this to be true [13] and so the results presented here should be treated with caution.

For Ki67 and Collagen IV, the null hypothesis could be rejected but the small numbers of samples Ki67 differentiated means that the use of this antibody is questionable. However, it seems that Collagen IV can be used to differentiate cancer from EM but the subjective nature of this analysis means that further work is required. It could have been the case that the discrepancies noted above are due to pathological interpretation rather than any biochemical differences.

The investigation into using combinations of different antibody staining scores found that pathological clustering could be achieved even when the most significant antibodies were left out of the multivariate HCA model. However given the small sample sizes involved, future investigations will need to use an enlarged cohort before any significance can be attributed to the classifications observed in the previous chapter.

Future studies will also need to consider the use of automated analysis algorithms for IHC. This will counteract the subjectivity of the above chapter. However, other more objective technologies such as Infrared spectroscopy (IR) could be used as an alternative. The next section describes how IR was utilized in this research and presents a new and novel digital staining technique.

## 4.3 Infrared Spectroscopy

Infrared (IR) spectroscopy was also investigated in addition to using IHC analysis in the discrimination of Epithelial Misplacement (EM) from cancer. IR spectroscopy is a powerful analytical technique that probes the vibrational modes of molecules and has been used previously within the literature to differentiate a variety of pathological conditions [48], [50], [51], [86], [89], [105].

Much research has been devoted to its development as a diagnostic adjunct and this is clearly evident from the current literature [50], [86–88], [92], [94], [154], [155]. The medical interest within this technique over the past few years has been aided greatly by advancements in technology that have allowed for its use within a diverse range of situations.

Important developments in this field include the attachment of motorised stages to IR spectrometers and their capability to process Fourier transformed spectra [98]. These advancements have allowed for IR to be utilised for clinical purposes and the results presented within the following chapter demonstrate its potential use for the differentiation of EM from cancer.

Specifically, this chapter contains information on the advantages and disadvantages of probing particular histological components with IR. It will be this sections purpose to demonstrate that the collection of histology specific Infra Red (IR) spectra is beneficial when developing functions for the automatic discrimination of Epithelial Misplacement (EM) from cancer.

### 4.3.1 Mucinous compounds within colonic disease states

The benefit of IR is that it enables the correlation of histology with spectroscopy. This in turn allows the biochemistry that is associated with pathological features to be quantified as IR spectroscopy probes the vibrational modes of molecules. This enables classification models to be built as disease states undergo distinct biochemical pathways that are different to benign tissue. The spectroscopic differences between these can then be used in the automation of pathology [50].

To enable this, efficient data processing techniques are required. One method of classification, is to convert the data contained within IR chemical maps into images that are analogous to stained tissue sections [47], [92], [155], [156]. This allows for the spectral differences between diseased and healthy regions to be visualised [50]. Alternatively, discrimination can be achieved by using an initial classification step in the collection of histology specific spectra [46], [88], [122] and then using these within multivariate models.

In this instance, the benefits of probing particular histological components, such as mucinous compounds, with IR spectroscopy were investigated. Mucin was chosen because much medical research has been devoted to analysing the structural differences that exist for this compound between different pathologies of the colon [88], [109], [110], [157–160]. These studies have mostly used antibody annealing to determine the presence of different mucinous compounds and they have identified pathology specific mucin expression profiles. For instance, it has been found that within conventional colonic carcinomas there is an up-regulation of MUC1[161].

It has also been found that for mucinous carcinomas there is an up-regulation of both MUC2 and MUC5AC proteins [115]. These protein types are synthesised in the goblet cells of epithelial tissue. Their dominance within particular pathologies has led some researches to believe that this may be indicative of a problem with the TGF-β signalling pathway. This is because this system regulates the synthesis of the mucin core proteins in this cell type[110].

Particular mucins have also been identified to have a role within inflammation and therefore have been causally linked to the development of colorectal cancer[162]. Therefore, the structural conformation of the mucins associated with cases of cancer should be different to benign samples [160], [163], [164].

#### 4.3.1.1 IR mucin specific acquisition methodologies

In this instance, a multivariate model was made from mucinous IR spectra acquired from four cases of normal tissue and four cases of cancer. The spectra collected were used to determine if any conformational differences existed between the mucinous compounds derived from the pathology groups. It should be noted here that the malignant cases used were sourced from resections and therefore were representative of fully invasive cancer.

As the investigated pathologies existed on the opposite ends of the diagnostic spectrum they should be the most biochemically different. Therefore, if accurate differentiation can be achieved between these, greater credence can be afforded to the classification of adenomatous cases from cancer by the targeting of mucin by IR.

In this study, the effective targeting of mucin was brought about by an image registration methodology that enabled IR and Haematoxylin and Eosin (H&E) images to be co-registered. This was necessary as it was difficult to localise mucinous regions from within unstained white light images of tissue sections. These are taken before IR maps are acquired and it is evident from Figure 4.44 that the mucinous regions of samples blend in with their paraffin background.

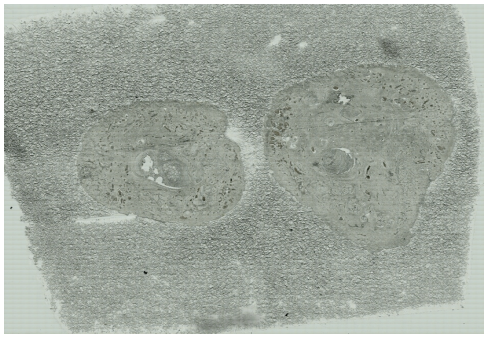


**Figure 4.44.** A representation of an unstained non-neoplastic case as imaged by the motorised stage of the Perkin Elmer Spotlight 400 Spectrometer (Massachusetts, USA).

However, mucin within H&E stained tissue sections is represented by a light pink colour, Figure 4.45. Therefore, the registration of consecutive stained sections to the unstained tissue will enable for this compounds accurate localisation. The use of consecutive levels from a tissue block is important as this enables for the overlapping of mucin pools, which allows regions to be identified that can be mapped with IR.

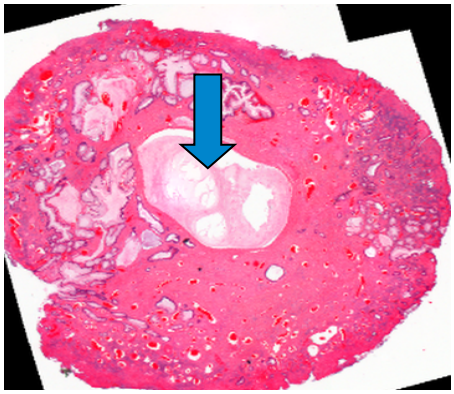


**Figure 4.45.** The mucin regions that are highlighted pink by H&E staining. The blue arrow indicates the mucinous region for this sample.

In this instance, registration was brought about by applying linear transformations to the H&E images [46]. The parameters of the rigid transformation were calculated from a Principal Axis Transformation (PAT) [38] and from optimising a Sum of Square Differences (SSD) [28] metric. For both the PAT and the SSD metric, the images used in the estimation of registration parameters were binary images. These were made from applying K means clustering to the first principal component computed for images [26].

The data matrix used in the binary conversion of images with PCA contained the spectral samples within its rows and the absorption intensities associated with these in its columns, Figure 3.3. The matrix design for PCA used in this part of the study was implemented throughout the rest of this chapter. Therefore, it should be noted that

any dataset used with PCA in the following chapter contained the sample numbers within its rows and the data values associated with them within its columns.

This is because for all of the IR and H&E images, the differences in intensity between the background slide and the tissue forefront were always described in this component. In this instance, two clusters were used with K means and this enabled for the tissue component from within classified images to be easily selected.

For the IR maps used in this registration the instrumental parameters used for acquisition were low and this enabled for fast imaging times. In this instance, images were acquired using a spectral resolution of 64$cm^{-1}$ wavenumbers and a spatial resolution of 50μm. These enabled for fast IR image acquisition times but still allowed for the associated spectra to be used in the differentiation of tissue from its background.

However, as noted within the 3D volume rendering section, Chapter 4.1, the linear registration of histology images is not appropriate because of the deformations that occur to tissue samples when they are being processed[147]. Therefore, on this occasion, precise co-registration did not take place but as the aim of the above procedure was to find and map large mucinous regions, the registration was not required to be absolutely accurate.

The final images transformed using the technique above is presented within Figure 4.46. The left hand side of these figures are the H&E images that have been deformed to look more like their IR image counterparts. The middle and end columns represent the respective binary masks made for H&E and IR images. It should be noted here that the H&E masks have been linearly transformed.

It can be seen from these figures that there is generally a good overlap between features. However, there are some obvious discrepancies between the two and this can be explained by the deformations that the individual sections underwent during histological processing.

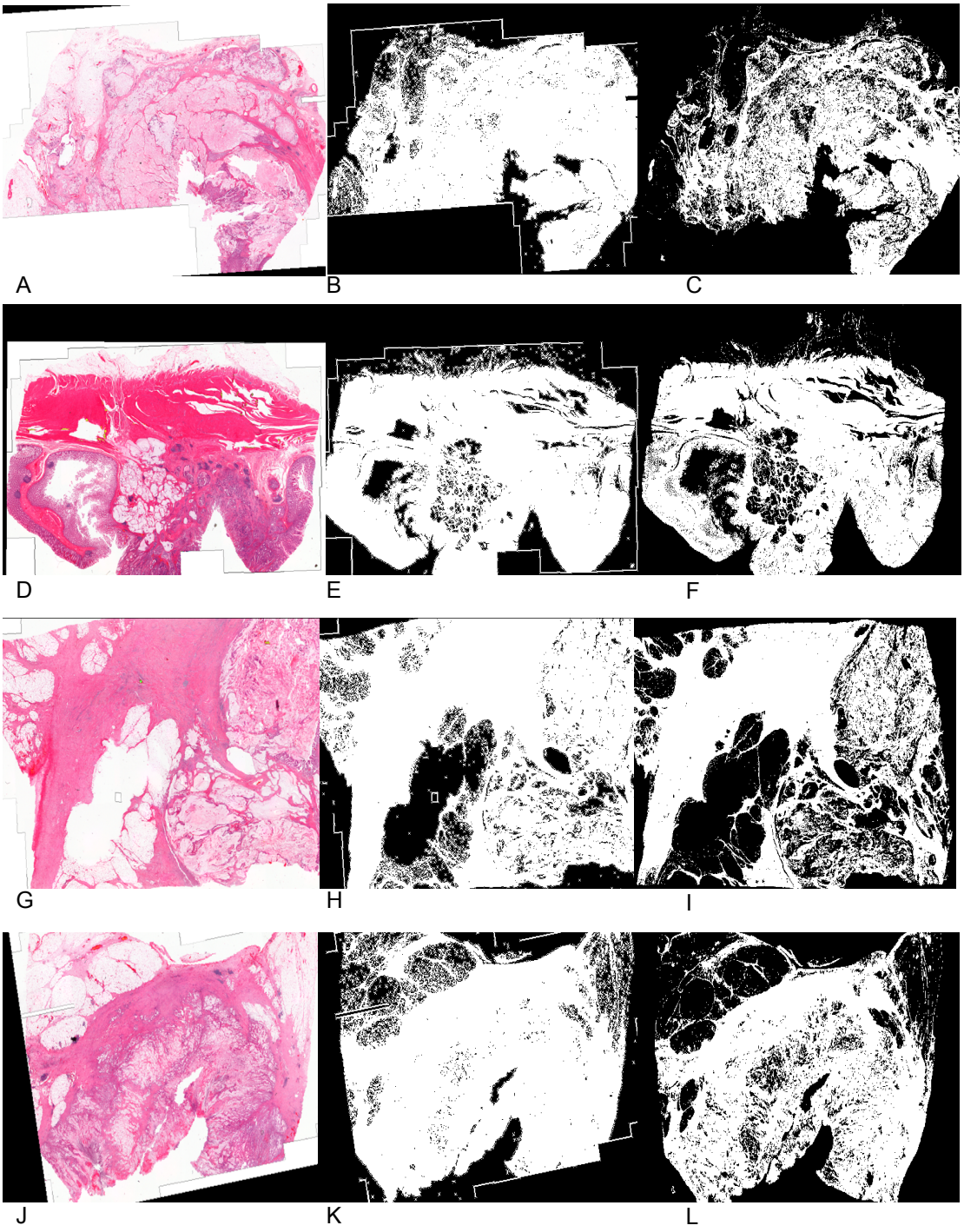



**Figure 4.46.** The linear registrations that were applied to cases of cancer so that the histology images more closely resembled their IR counterparts. Left column: The H&E images of the cancer cases with the mucinous regions clearly visible. Middle column: The Boolean transformation of the images found on the left hand side. These images have been rigidly transformed so that they more closely resemble their corresponding IR images. Right column: The binarised IR images made from carrying out PCA and K means clustering on low spectral and spatial resolution IR maps.

4.3.1.2. Selection of IR spectra of mucin

From the images presented in Figure 4.46 the location of the mucin regions to map with IR were determined. However, it is also obvious from these figures that the mucin regions are infiltrated with other histological components, such as the submucosa and epithelial glands. Therefore, a method of collecting spectra only representative of mucin was required.

There are many ways in which spectra of this sort can be collected. For instance, IR mucin maps could be entered into a classification model as a whole [94], point spectra of mucinous regions could be taken [87] or the spectra associated with mucin regions could be segmented from IR images and used further [88]. The first two options will add undesirable variation to the recorded dataset and therefore the latter option was pursued.

For the segmentation of mucin to occur, 0.3cm by 0.3 cm IR maps were acquired using a modest spectral and spatial resolution of 6$cm^{-1}$ wavenumbers and 6.25µm respectively. For these images, two measurements were coadded at every pixel, as the mucinous regions of samples were quite large. These parameters have been used elsewhere in the literature and have enabled for the accurate differentiation of disease states [87], [88]. The use of a higher number of coadded scans would have meant that the image acquisition times would not be feasible and it would also mean that only small mucinous regions could be acquired with IR.

The higher resolution IR maps of particular mucinous regions were analysed in multiple ways so that the segmentation of mucin could be brought about. For instance, initial investigations used only the amide peak regions from the IR spectra within a multivariate statistical model. It was hoped that this would lead to the effective classification of mucin as these compounds are essentially proteins that are thought to have different peak positions to the rest of the sample [88].

Similar methodologies have been tried elsewhere in the literature to some success and it was hoped that this could be repeated here [165]. However, it was determined that mucin could not be efficiently segmented in this way.

Therefore, the possibility of using the spectral information contained within a target image to train a statistical model was investigated. To do this, the loading of a PCA component (Figure 4.47), where the mucin was obviously defined, was used to identify the location of mucinous regions from within other samples.

In this instance, performing PCA on all pre-processed chemical maps identified the above loading. On this occasion, pre-processing involved taking the acquired spectra's second derivative by using a third order 15 point Savitzky-Golay filter; removing the

regions associated with paraffin; normalising the spectra with respect to their magnitude and by mean centring them.

The reason for taking second derivative spectra was so that shoulder peaks could be distinguished and because this technique removes baseline artefacts. The decision behind the removal of paraffin peaks from this analysis will be discussed later. In this instance, a suitable loading for the cancer cases could be found and is presented in Figure 4.47. The main peaks identified within this PCA load can be found at 1662, 1527, 1098 and 1011cm$^{-1}$ wavenumbers.

The peaks at 1011 and 1098cm$^{-1}$ wavenumbers, which are highlighted in this figure, relate to the carbohydrate region of the infrared spectrum. Specifically, they relate to the molecular vibrations of C-O bonds[166]. This should be expected as the carbohydrate structure of mucin would have strong absorbance in this region[110].

The other peaks identified, present at 1527 and 1662cm$^{-1}$ wavenumbers, relate to the amide region of the infrared spectrum and these peaks have been specifically identified within the literature as being related to mucin [92]. Therefore, confidence can be asserted that the mucinous regions of tissue samples have been identified by this component. Further credence in the above can be found in the fact that within the predicted scores the mucinous compound is consistently represented, Figure 4.48.

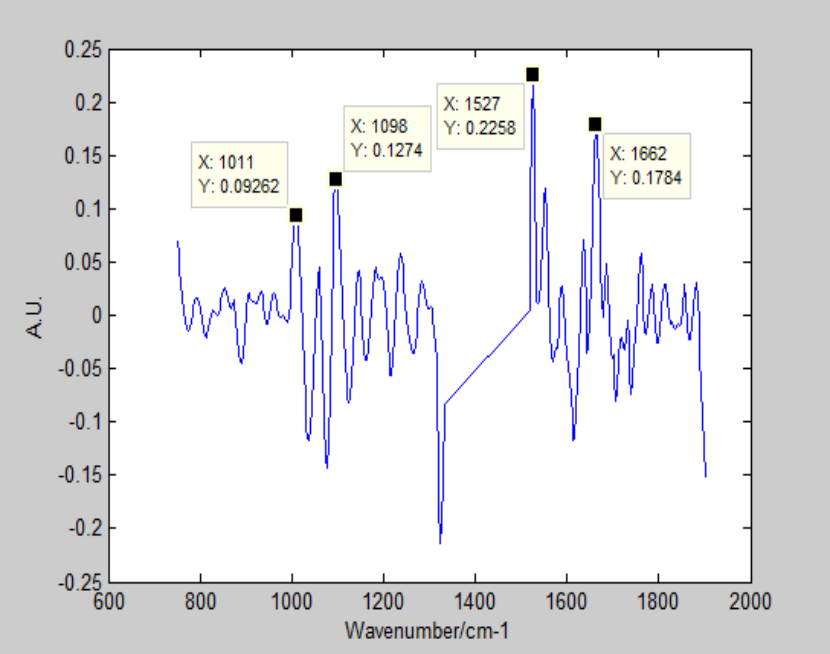



**Figure 4.47.** The loadings of the target image's second PCA component. This was used further in order for the mucinous regions of other cancerous tissue samples to be segmented.

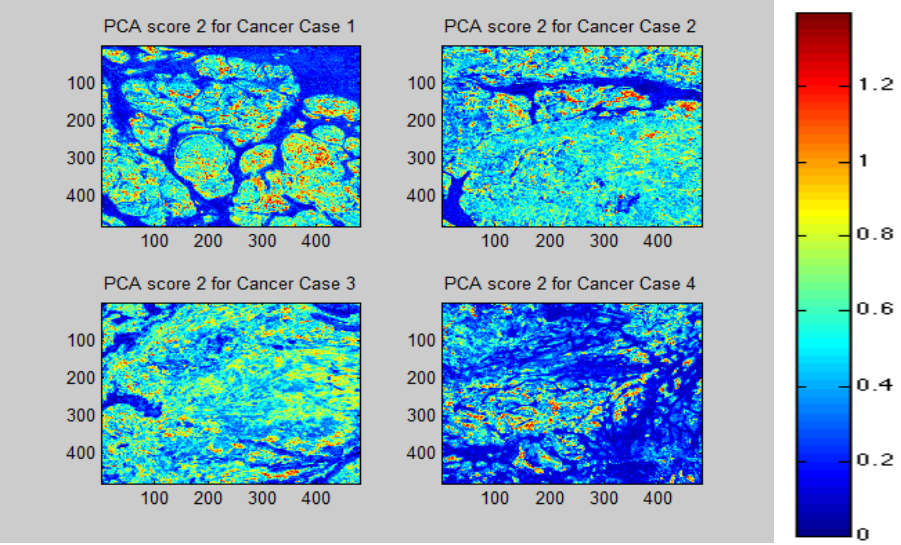



**Figure 4.48.** The predicted location of mucinous components from within images using the loadings presented in Figure 4.47. The image from which the mucinous loadings are derived is presented in the top left hand corner.

Ideally, a single loading should be used in the prediction of the location of mucin from within images of both cancer and normal tissue. However, this segmentation procedure was performed for cancer and normal samples independently. This is because it was the spectroscopic benefits of probing mucin with IR that were being investigated. Therefore, the use of this methodology was justified. There were similar peaks found within the loading made from the normal target image but an additional beta protein conformational turn peak was present at 1674cm$^{-1}$ wavenumbers [167].

The final segmentation results for both the normal and malignant cases are presented within Figure 4.49 and Figure 4.50. These were made from using the 95% confidence limits of the D Statistic with a training set of mucin specific scores. It is evident from Figure 4.49 & 4 that this approach was successful and that it allowed for the automated collection of mucin spectra from within the measured samples.

To demonstrate the accuracy of this technique, the H&E images for the normal cases can be found in Figure 4.51. It can be seen from this figure that the mucin has been accurately segmented and further evidence of this comes from when the IR and standard histology images are overlaid, Figure 4.52.

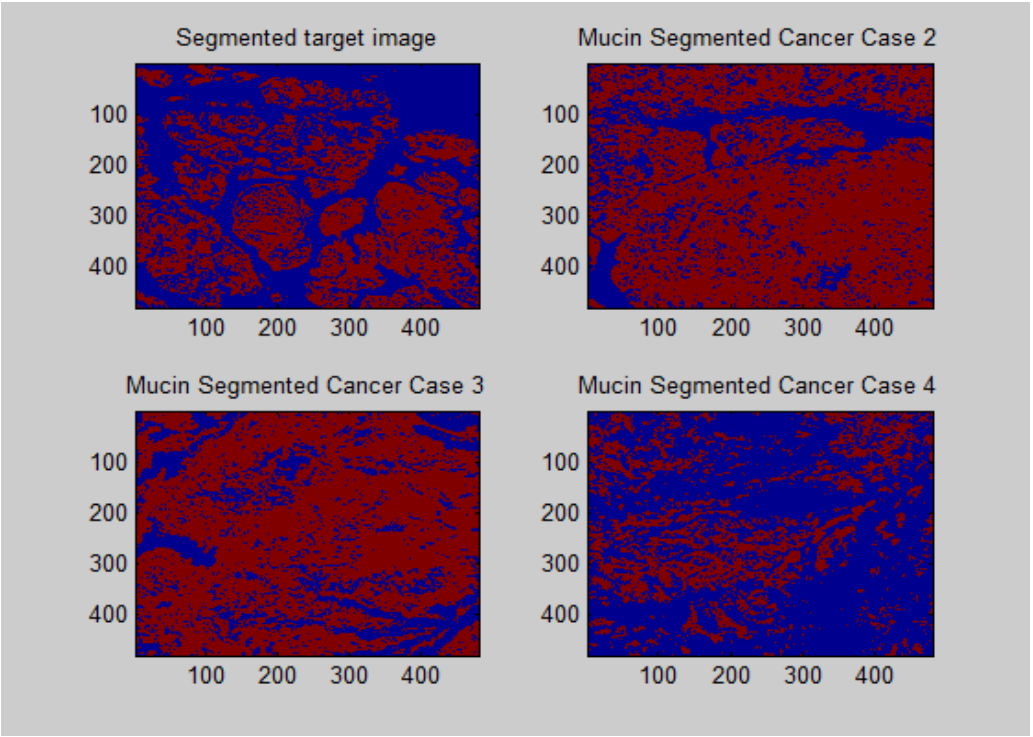



**Figure 4.49.** The segmentation of mucin from within IR images of cancer samples. The target image used in this instance, that enabled the segmentation of mucin to occur, is presented in the top left.

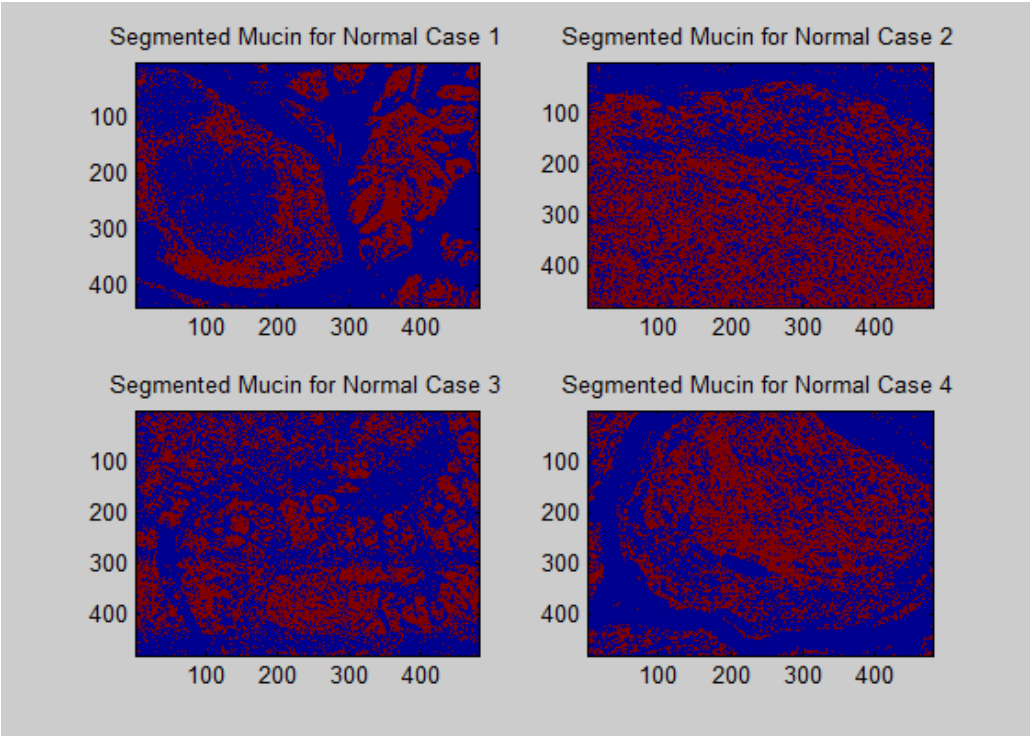



**Figure 4.50.** The segmentation of mucin components from within IR maps of normal biopsies. The target image used in this instance, that enabled the segmentation of mucin to occur, is presented in the top left.

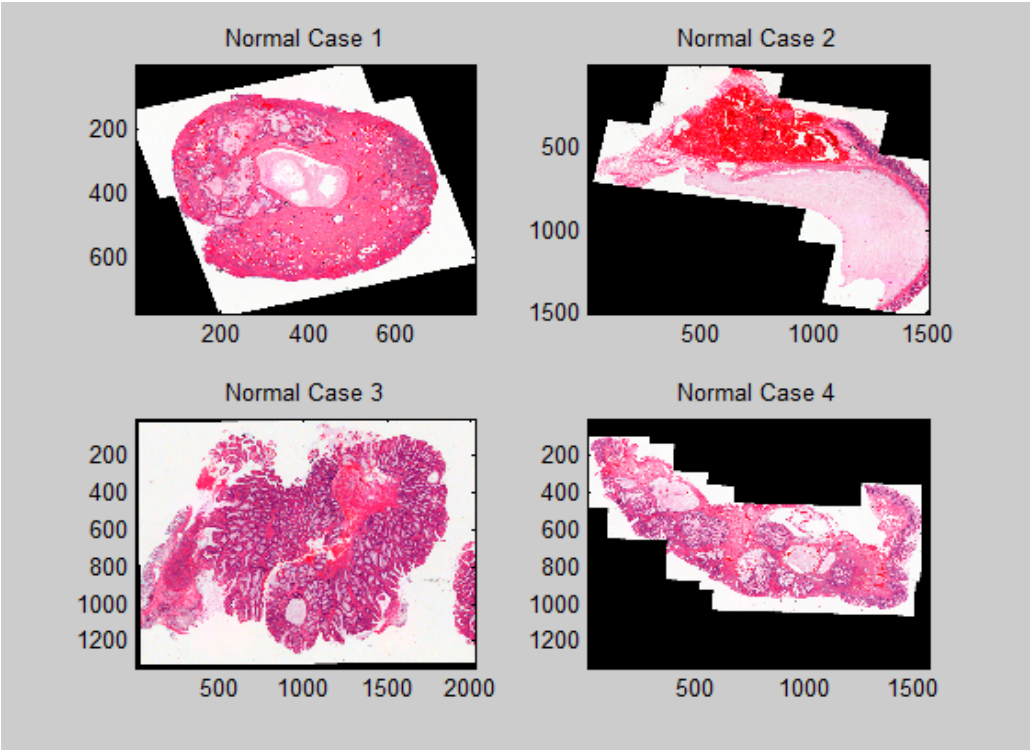



**Figure 4.51.** The corresponding linearly registered H&E images for the segmented images presented in Figure 4.50.

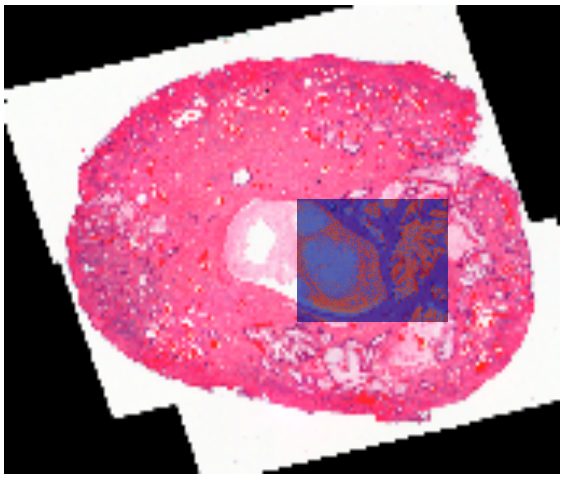


**Figure 4.52.** The segmented IR map for one of the normal cases overlaid on its H&E counterpart.

Although, this segmentation methodology was highly accurate it was not without its problems. For instance, in one sample, it appeared that there were problems associated with the accurate segmentation of mucin. This is evident from the bottom right hand corner of Figure 4.49. In this figure the mucin regions are sparse relative to where the mucin can be seen within the H&E of this sample, Figure 4.46 G.

However, the random selection of pixels from where it was thought there should be mucin revealed epithelial like spectra were present (these spectra are coloured red in Figure 4.53). Confidence can be asserted that these are epithelial spectra as there are nucleic acid peaks present at 1077 and 1242cm$^{-1}$ wavenumbers. Within the literature it has been noted that these peaks are representative of phosphate bond stretching [85], [154].

The closer inspection of Figure 4.46 G reveals that there was an infiltration of glandular tissue within the measured sample. Therefore, contrary to first appearances, this segmentation method may have worked quite well. For those spectra that classified as expected it can be seen from randomly collected spectra, those coloured in green in Figure 4.53, that the well defined mucin peak at 1044cm$^{-1}$ wavenumbers was present [88]. This means that mucin has indeed been sampled using the above approach and the D Statistic is rejecting other histological components.

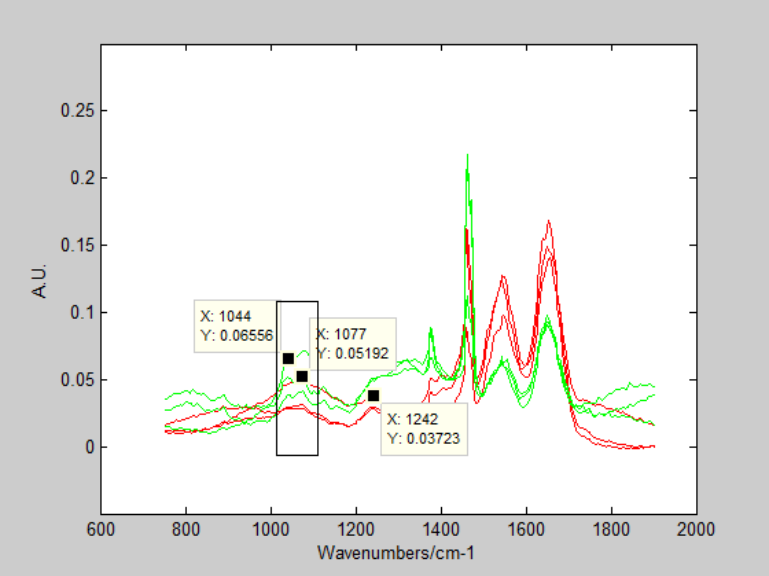



**Figure 4.53.** The infrared spectra of selected pixels from the bottom left image of Figure 4.49. The red spectra are pixels that have been rejected by the D Statistic and the green spectra are those that fall within the 95% confidence limits of the training set.

4.3.1.3 Pre-processing of mucin IR spectra

Using the segmentation technique presented above around four hundred thousand spectra were collected from the normal maps and the same number was collected from the cancerous samples. These spectra were pre-processed by taking their second derivative, removing their paraffin peaks, vector normalising and finally mean centred them.

The removal of paraffin before classification, and segmentation, was deemed necessary because early analysis suggested that the variance in these peaks were dominating the PCA loads. This is evident from Figure 4.54 that is representative of a PCA fed Linear Discriminant Analysis (LDA) model that was built when the paraffin peaks were left within unprocessed spectra.

Within the bottom right of Figure 4.54 it can be seen that the largest LDA weight corresponds to the second principal component. The weights of LDA correspond to PCA components that enable the biggest discrimination between pathology groups to occur. As the largest weight corresponded to a component where the paraffin peak was being so strongly described it was determined that no biochemical discrimination was occurring within this model.

Therefore, a method of accounting for paraffin was required. Within the literature there are many developments that have enabled for the removal of spectral contributions from paraffin. A few of these options, which can be either digital [88], [101], [102], [104] or chemical in nature[105], [168], were explored in the following analyses.

When using chemical means, such as Xylene to dewax samples, caution should be exercised, as the process is not always 100% complete. It has also been noted within the literature to cause chemical damage to tissue samples. For instance, it is known to affect lipid molecules that are contained within the membranes of tissue sections [89]. This might be problematic for colonic tissue samples as it has been shown within the literature that the phospholipids associated with colorectal cancer might be important in its differentiation from adenomatous and normal tissue samples [94].

It is because the chemical effect of dewaxing reagents on tissue samples is not completely defined, the literature has developed digital means to remove paraffin. Methods for the digital removal of paraffin include Independent Component Analysis (ICA) with non negative least squares fitting and the Extended Multiplicative Scatter Correction (EMSC) algorithm [102], [104]. The ICA approach was developed with Raman spectra in mind and has been noted within the literature as not currently being appropriate for use with infrared spectroscopy[104].

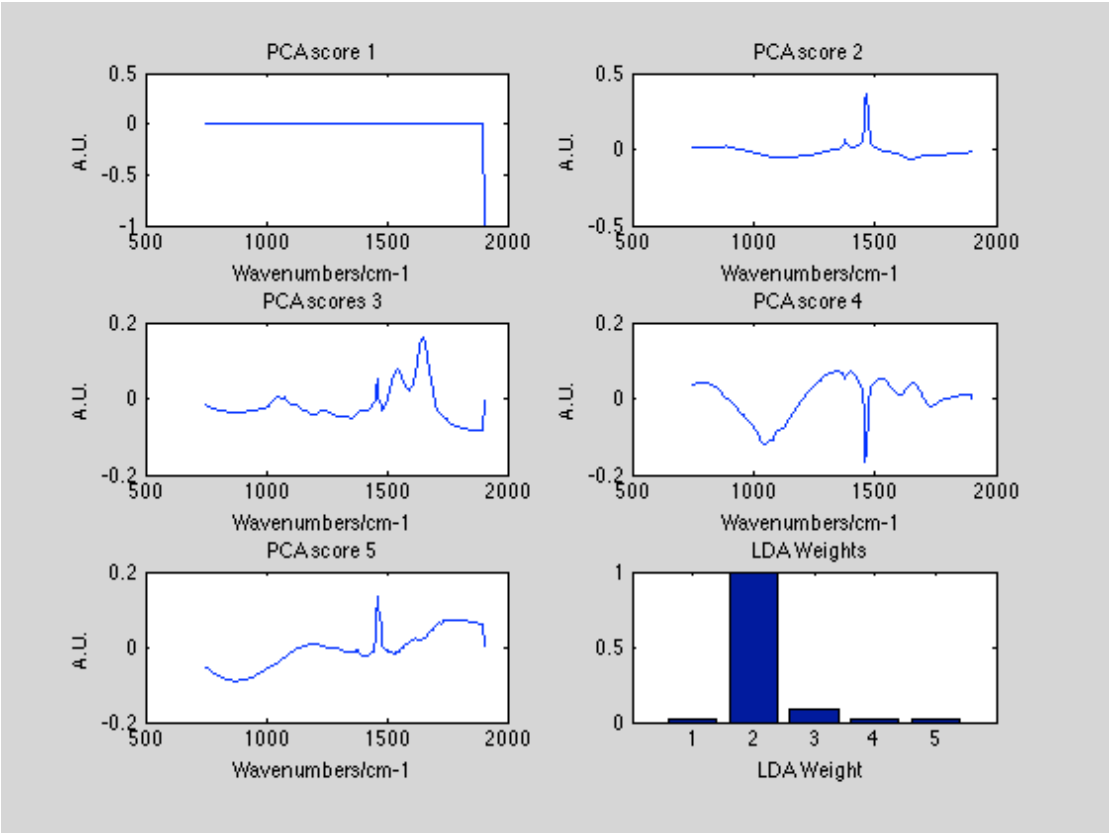



**Figure 4.54.** The dominance of paraffin within the PCA loads of a model where the spectral variance between mucin was being analysed. In this instance, no spectral preprocessing was performed. The first five plots relate to the loadings of the PCA model and the last bar chart corresponds to the LDA loads.

Therefore, the EMSC algorithm was used and has been used extensively within the literature to improve the visualisation of histological features within paraffin embedded IR chemical maps [88], [102]. However, the power of infrared spectroscopy comes from its use in the differentiation of disease states and therefore requires high quality spectra.

In this instance, it was noted that EMSC added undesired volatility to the recorded spectra. Therefore, processing in this way does not enable the development of reliable classification models. The artefacts that EMSC introduces into spectra are evident from the bottom of Figure 4.55. As such, the removal of paraffin peaks from analysis was the only viable option that negated some of the problems identified above.

The paraffin region removed from the recorded spectra was determined from visualising the mean of an area of an image that exclusively contained this compound, Figure 4.56. In this instance, the recorded spectra between 1503 and 1350cm$^{-1}$ wavenumbers were removed from all spectra.

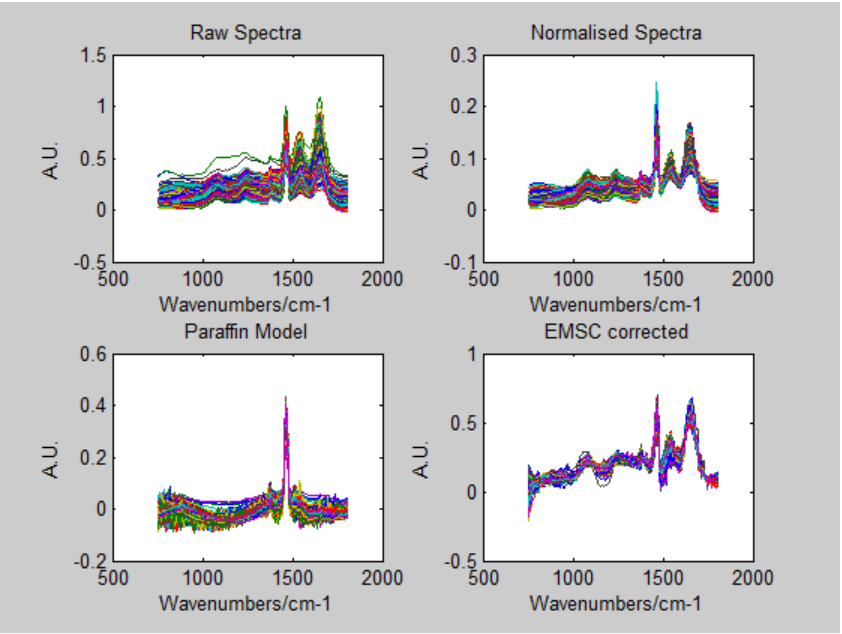



**Figure 4.55.** The anomalies that the Extended Multiplicative Scatter Correction (EMSC) algorithm introduced into spectra. Top left: The IR spectra that were recorded by the Perkin Elmer Spotlight 400 spectrometer. Top Right: The normalized spectra of those presented in the top left. Bottom left: Normalized paraffin spectra that have been abstracted from within infrared images. The dominant peaks of the paraffin spectra are obviously present within the tissue spectra. Bottom Right: The EMSC corrected spectra where paraffin was used as the interference spectrum.

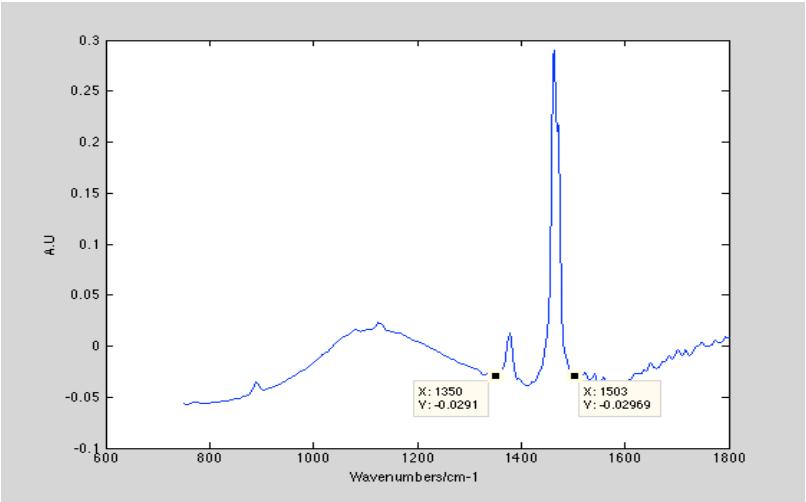



**Figure 4.56.** The mean spectra of paraffin taken from an IR map.

Other problems associated with the recorded mucin dataset was that there were large background artefacts present within the recorded spectra. Separate PCA fed LDA

analysis showed that these physical phenomena were largely significant within classificatory models.

From the literature it has been determined that these artefacts are a consequence of using the transflection mode of the IR system [100], [101]. Data collected in this way will contain anomalies because the IR optical path travels through many different entities, each of which will have different refractive indices. This causes the IR energy to deviate from its path and results in some of the unabsorbed energy not reaching the detector.

EMSC has been used with an appropriate reference and interference [100] spectra in the literature to correct for this but as noted in Figure 4.55 this method is not fit for purpose. It is true that the parameters of the EMSC algorithm could be adjusted and an iterative approach could be taken in the derivation of a suitable reference spectrum. However, these parameters are likely to be specific to the acquired data and would mean that any model made from these might not be generalizable to all samples.

Therefore, in an attempt to remove these physical effects a second derivative 15 point third order Savitzky-Golay filter was used. Although, it should be noted that this technique will only correct for sinusoidal baseline phenomenon. It will not correct for distorted peak positions, that are most likely to occur around the amide I region of the spectrum [101]. However, a more thorough investigation into the effects of scattering is required and alternatives to transflectance, such as using transmission spectroscopy with calcium fluoride slides, should be considered.

It was thought that second derivative spectroscopy would be useful as it resolves the shoulder peaks of IR spectroscopy. However, as noted recently within the literature, the use of derivative spectra might not be appropriate for the identification of robust biomarkers [140]. This is because Savitzky-Golay filtering will smooth spectra and will ultimately cause the distortion of peak shapes and positions [140].

In this instance, the quality of the recorded spectra was also scrutinized. Those spectra that were two standard deviations above or below the mean of the recorded datas integrated amide peaks were excluded from further analysis. This enabled for low intensity and saturated spectra to be removed from the modeled dataset and prevented them from having an effect on the final classification results.

#### 4.3.1.4 Classification of pathologies by mucin specific IR spectroscopy

The pre-processing identified above enabled for the potential use of mucin as a discriminatory agent to be evaluated. In this instance, a PCA fed LDA model was used in the production of discriminatory functions. The 20 components used with LDA were selected by two way ANalysis Of VAriance (ANOVA) as being the most discriminatory.

LDA is a technique often used with IR spectroscopy and has a proven track record within the current literature [90], [94], [95], [143], [169]. Therefore, it was hoped that a model could be developed that would facilitate the discrimination of normal mucin from cancer. However, upon subsequent analysis of the PCA loadings with their LDA weights it was determined that no discriminatory biochemistry was being described by the model, Figure 4.57.

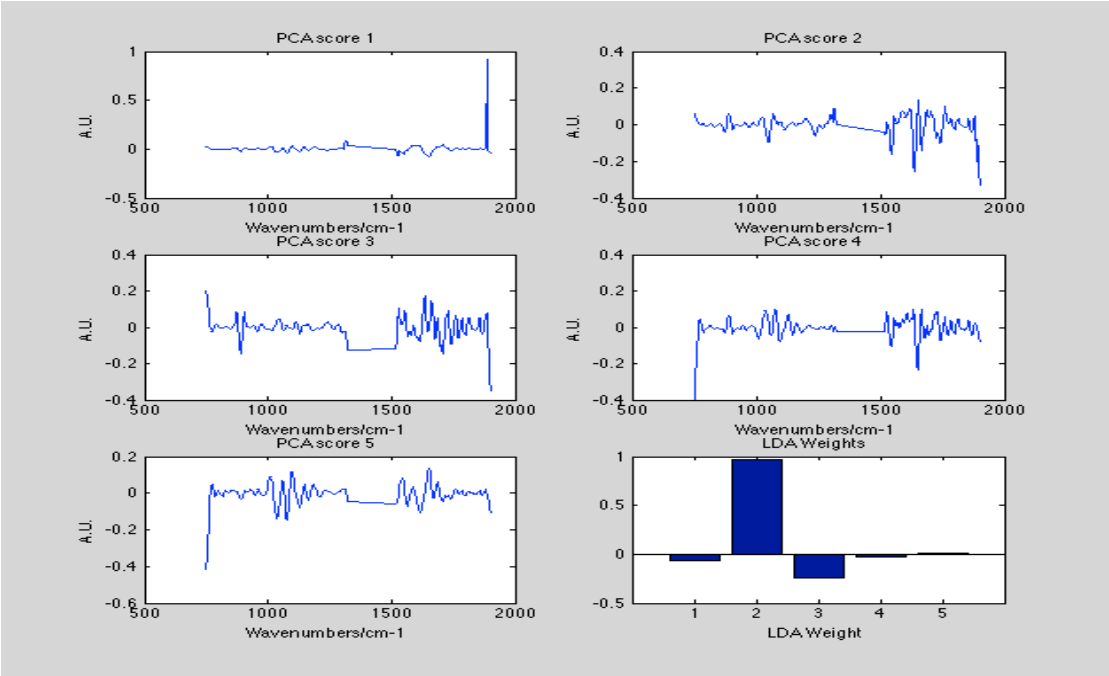



**Figure 4.57.** The most significant principal components as determined by ANOVA and their associated LDA weights.

It can be noted from the bottom left of Figure 4.57 that the second principal component was the most dominant in the differentiation of the disease states. Within the PCA loadings associated to this weight it can be noted that the most significant peak relates to an artifact introduced by the Savitzky-Golay second derivative filter. It could be argued that this region of the spectra be removed and that the analysis be re-run. However, as a large cohort of spectra was used in this instance, it is likely that if there were any significant biochemical differences between pathologies it would be highlighted using the recorded spectra.

The lack of differentiation between pathology groups when using mucinous IR spectra throws doubt on its clinical use. This is because the colonic samples measured here were on the opposite ends of the pathological spectrum. Therefore, it is unlikely that any success would be had in the differentiation of adenomatous polyps from cancer when targeting this compound.

#### 4.3.1.5 Differentiation of adenomatous mucin from cancer

To fully explore the appropriateness of targeting adenomatous mucin with IR an investigation was conducted, whereby, four IR images of mucinous regions of polyps and cancerous tissue were acquired. In this study it was decided that chemical dewaxing should be implemented, as it is obvious from the above investigations that paraffin represented a particular problem. Simply removing the paraffin peaks may have been too coarse as important biochemical information may have been lost.

In this instance, the removal of wax enabled for the easy identification of mucinous areas and therefore meant a linear registration was not needed. However, the stitching together of H&E images was still necessary as it enabled confidence to be asserted that the correct histological features were being analysed.

For the sampled regions, the segmentation procedure developed earlier was used again in the collection of spectra relating to mucin. The results of this segmentation procedure for the measured cases are presented within Figure 4.58. It can be noted that the left four images of Figure 4.58 are highly comparable to those presented in Figure 4.49. However, for the adenomas it can be noted that the segmentation procedure was extremely efficient in two cases and not so reliable in the others.

In the cases where the mucin segmentation was poor it can be seen that this was either due to spectral anomalies or because there was no mucin present. Within the image found second from the bottom right hand corner of Figure 4.58 it can be seen that the mucin regions segmented are sparse in appearance. When viewing random spectra derived from the regions where mucin should be expected it is obvious that the spectra associated with the misclassified regions is of a poor quality (Figure 4.59).

When the second derivative of the spectra presented in Figure 4.59 was taken its associated noise would have prevented it from classifying properly. For the other poorly segmented sample, it can be seen that there are no mucin pools present within the region acquired by IR after histological processing (Figure 4.60). It should be noted here that the mapped region was the only prominent mucinous pool that could be found within the sample.

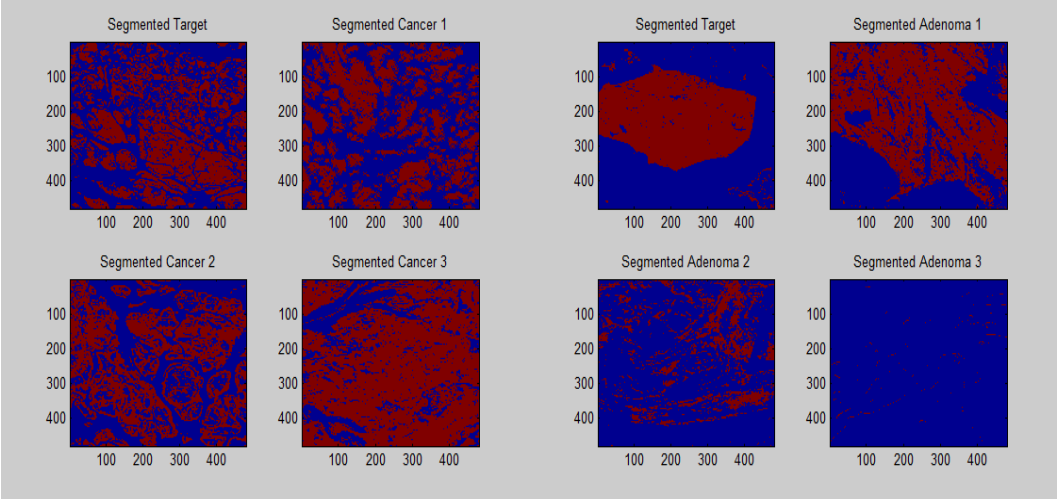



**Figure 4.58.** The segmentation of mucin from within IR images of cancer (left four figures) and from within adenomas (right four images). H&E images for the adenomatous samples are provided within Appendix B.

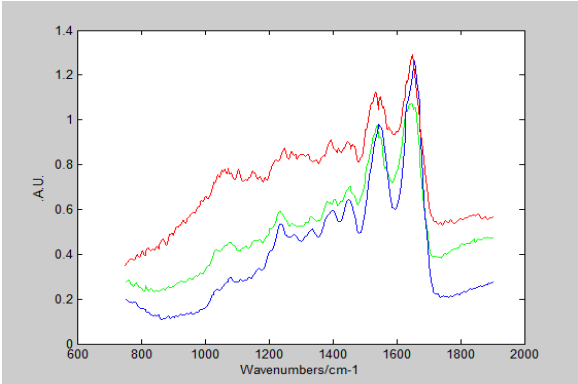



**Figure 4.59.** Random spectra from misclassified regions of the segmented image presented in the image second from the bottom right of Figure 4.58.

Nevertheless, mucin spectra representative of adenomatous tissue could still be derived and its use was felt to be valid within a multivariate model. In this instance there were over four hundred thousand cancer spectra collected and around two hundred thousand spectra were derived from the adenomatous samples. The difference between these sample sizes can be explained by the difficulties presented above in the accurate identification of mucin.

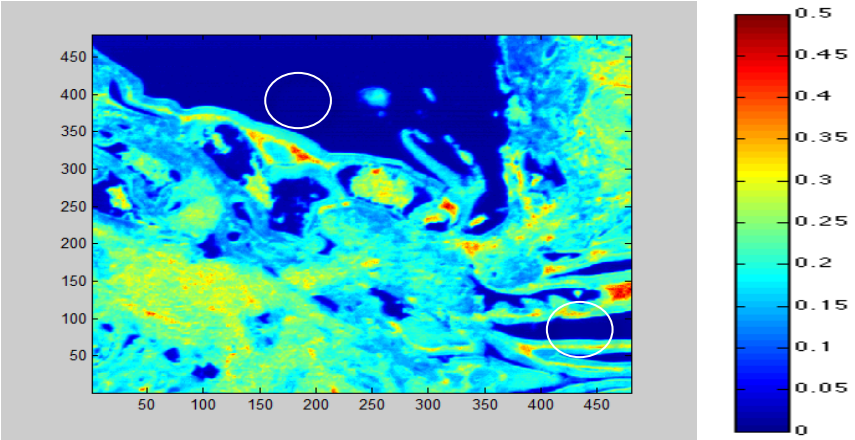



**Figure 4.60.** The lack of mucinous pools within an adenomatous IR map. This is a PCA score for the tissue sample and since the expected mucin regions have the same value as the background slide we can be confident that no mucin was present. Regions of the background slide and the mucin pools are highlighted by white circles.

Again, spectra were processed using a 15 point third order Savitzky-Golay filter before they were normalized and mean centered. However, it should be noted that the paraffin peaks were left in in this analysis as the tissue samples had been previously chemically dewaxed.

The further implementation of PCA and ANOVA enabled the selection of 20 significant principal components and these were included within a LDA model. The PCA loadings and LDA weights for the most significant principal components are presented within Figure 4.61.

It can be seen that the most discriminatory PCA components, as determined by LDA, mainly corresponded to spectral artifacts that was probably introduced by the Savitzky-Golay filter. This means that the 85% and 84% spectral classification achieved for the mucinous spectra using this model should be treated with caution. However, the use of this filter was justified because of the presence of baseline phenomenon.

The physical phenomenon that was mainly causing the differentiation of disease states is clearly evident in principal components one and four. This is because of the presence of uncharacteristically dominant peaks at the high-energy regions of the infrared spectrum. These components have the largest LDA weights, bottom right of Figure 4.61, and they are probably describing differences between spectra that have this spectral defect and those that do not.

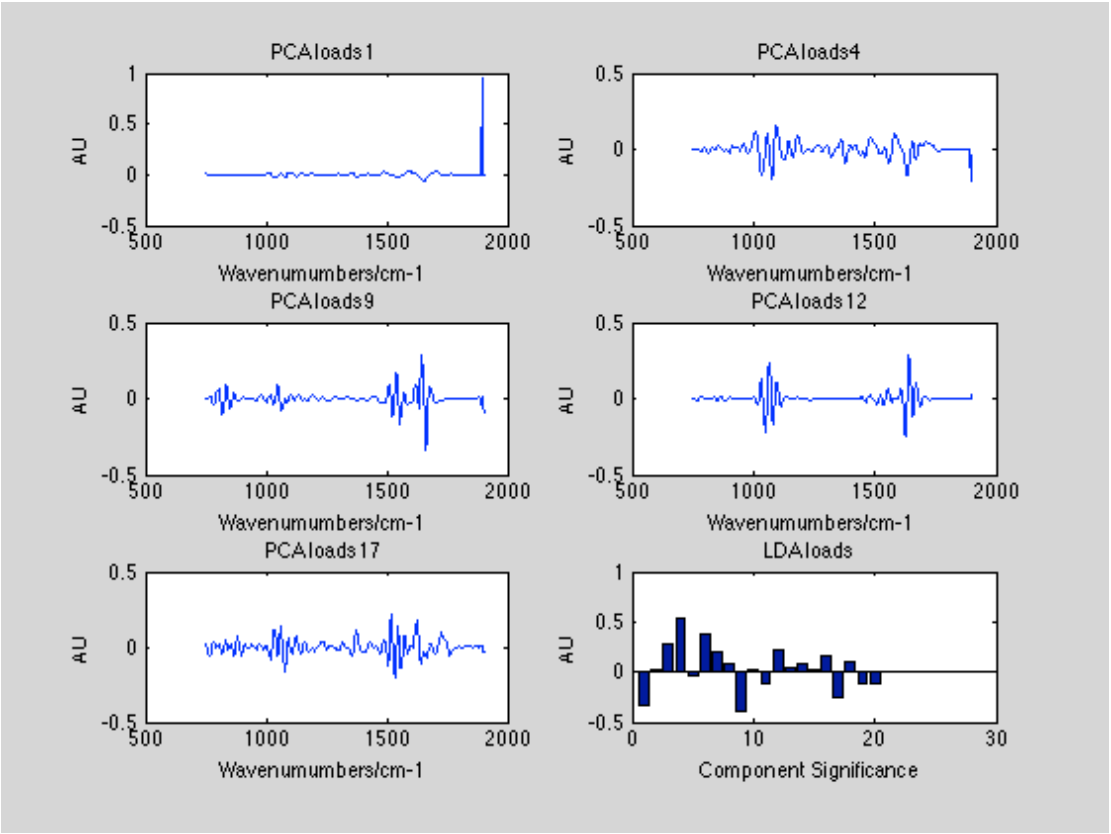



**Figure 4.61.** The most significant PCA components as determined from the first five largest LDA weights. They describe the variation that exists between mucinous IR spectra that was derived from adenomatous and cancerous tissue samples. In this instance, the most discriminatory principal components chosen were the loadings corresponding to principal components one, four, nine, twelve and seventeen.

The significance of principal component nine is on par with the first and it maybe that this is describing some discriminatory biochemistry. For this component, differences between the pathologies amide I region could be seen [170]. For cancer, there was a significant peak present at 1644cm$^{-1}$ wavenumbers and for the adenomatous tissue an important peak was present at 1662cm$^{-1}$ wavenumbers.

A review of the current literature would suggest that chemical dewaxing reagents can cause changes in peak intensities and positions within this region of the IR spectra [105]. Therefore, any changes in this part of the IR spectrum should be treated with caution. Further problems associated with discrimination around the amide I region of the IR spectra can be related to scattering [100]. Although, other peaks are present, they are not as significant as the protein peaks. Therefore, caution should again be taken before any significance is attributed to them.

For the other components, twelve and seventeen, it can be seen that the main spectral differences relate to the protein and carbohydrate regions of the infrared spectrum. However, it should be noted that the significance of the twelfth component in the LDA weights was much lower than the rest. Therefore, no further analysis was conducted on this component.

For the seventeenth component a significant fatty acid peak was present at 1725$cm^{-1}$ wavenumbers [139]. However, caution should be taken when attributing any significance to this peak because chemical dewaxing was used [168]. Further problems with this load include the appearance of multiple peaks. This might be reflective of the fact that second derivative spectroscopy was used but also could be due to noise.

For the analysis presented above, it could be argued that the IR spectra be truncated in regions where spectral defects are obvious and that the model be rerun. However, the truncation point will be hard to accurately identify and any results subsequent to this might still contain anomalies that are not obvious. Also, by carrying out this sort of analysis it might be the case that the classification of data is biased towards the training dataset.

Another limiting factor associated with the above analysis was that the collected mucin spectra were representative of all the chemical components contained within the sampled mucinous pools. As noted by Walsh et al and Travo et al [88], [110] the pools contained within colonic tissue are a mixture of many different molecular components. Therefore, although confidence could be asserted that in this research mucinous spectra were being acquired, it was highly likely that these were contaminated with other molecular signals.

This would have impacted the multivariate classification models that this research implemented. For instance, these additional components could have masked important diagnostic peaks that would have otherwise been capable of bringing about the differentiation of EM from cancer. Therefore, in the future, other data exploration methods should be used when considering the analysis of mucin.

For example, a least squares solution could be used with reference mucin signals in the determination of the contribution that they were making to the recorded spectra. Although, caution should be taken when using this approach with the data that was acquired by this research. This is because a low number of scans were used to acquire the IR spectra of the mucinous pools. Therefore, it will be highly likely that spectral artefacts will impact the fitting coefficients found from the above analysis.

In this investigation, an attempt was made at acquiring spectra representative of pure mucin compounds by dissolving Sigma Aldrich Bovine Mucin Type II into solution and

drop plating this onto low-e slides. The outer rings of these drops were then measured using the point spectra mode of the Perkin Elmer Spotlight 400 Spectrometer with a spectral resolution of 4$cm^{-1}$.

However, this investigation found it extremely hard to derive unsaturated mucin spectra, i.e. the only signals acquired were due to intense molecular concentrations. This research attributed this saturation due to the lack of mucin dissolution. As a consequence, in this research, it was not possible to assess whether components other than mucin were impacting the recorded spectra. However, it should be noted that experiments carried out by Travo et al found that the mucin signals they recorded did not deviate too far from the reference mucin spectra that they acquired.

Therefore, other histological targets were sought. This is because chemical and physical artifacts dominated the above PCA loadings and because viable mucin glands were not present within all adenomatous samples. Consequently, no further analysis was conducted on the collected mucinous spectra. Although, future investigations should review the use of targeting this biological entity when more sophisticated digital paraffin removal methods become available.

### 4.3.2 Epithelial specific spectra

The next target chosen for analysis with IR spectroscopy was the epithelium. This is because it is abundantly present within colonic tissue sections and is one of the histological entities that are routinely used by pathologists in the differentiation of disease states.

For this study, it was decided that the IR images would be acquired on tissue samples that were still embedded in their paraffin encasing. This is because of problems associated with chemical dewaxing [88], [104], [105] and because of the noise EMSC introduced.

The use of chemical dewaxing agents are especially problematic for colonic pathologies because these reagents affect membranous phospholipids [92], [94], [168]. The chemical composition of the membranes of both cancer and adenomatous tissue has been noted as being important in their characterization within the literature [94]. Although, it should be noted that before tissue samples are embedded in paraffin, they may in some cases be washed in xylene. However, the further application of this reagent will only add to any observed alterations.

Therefore, in this instance, IR maps were recorded on paraffin embedded tissue sections and the paraffin peaks were removed from any subsequent analysis, after second derivative filtering. For this study, images that were acquired were done so

using a spatial resolution of 25μm and a spectral resolution of 6cm$^{-1}$ wavenumbers. In total, 11 chemical maps were acquired for both cases of EM and cancer.

4.3.2.1 Derivation of epithelial specific training models

To enable the spectroscopic characteristics of the epithelium to be probed an appropriate segmentation methodology had to be developed. In the first instance, a similar segmentation technique as used with the mucin IR maps was implemented. This involved using the recorded spectra in the determination of an epithelial specific PCA component. If such a component could be found then the loadings associated with it could be used further in the identification of the epithelium from within other samples.

For all measured IR maps PCA was performed and the resultant score maps were scrutinized so that an epithelial component could be found. Before analysis, the IR spectra that related to the background slide and paraffin were removed so that non-informative spectra were left out of the dataset. This was achieved by carrying out PCA and then using K means clustering with the first principal component.

Subsequent to this, the remaining spectra within the IR maps were processed by using a 15 point third order Savitzky-Golay filter. The relatively high window size used was because only two scans were co added when the IR data was acquired.

Relative to the rest of the literature this window size sits somewhere in between what has been used previously with tissue samples [48], [171], [172]. However, caution should be exercised when using filter sizes that are too high, as these will affect bandwidths and peak positions. This can lead to erroneous conclusions being drawn about any spectroscopic differences that exist between pathology groups[140]. Although, it should be noted that within the literature there is a wealth of research that has utilised derivative spectra in the differentiation of disease states [88],[48], [87], [89], [90].

The IR PCA score map where the epithelium was most identifiable is presented within Figure 4.62. However, it should be noted that the submucosa was being identified here. This is because the submucosal features identified in Figure 4.62 are not only apparent in the PCA score but also in its loadings, Figure 4.63.

It can be seen from Figure 4.63 that significant peaks are found at 1221, 1539, 1614 and at 1686cm$^{-1}$ wavenumbers. The peaks at 1539 and 1686cm$^{-1}$ wavenumbers were related to the β sheets of proteins[88], [90], [170]. This is because the literature suggests that collagen, which is a major constituent of the submucosa, has β sheet like conformations [173]. However, the peaks attributed to submucosal proteins are purely speculative. This is because no exact correspondence with the literature exists. For the

other peaks, no other logical assignments could be made and this maybe representative of noise.

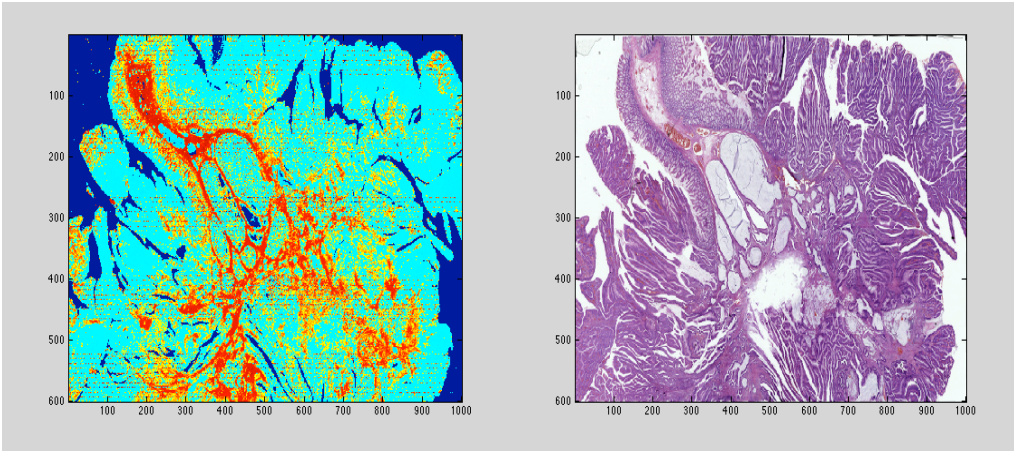


**Figure 4.62.** An image demonstrating the comparable features that exist between a PCA score map and an H&E section that was made for the presented sample. Left: the thresholded IR PCA score map. The red colour relates to the submucosal tissue and the blue corresponds to everything else. It should be noted that the colour scale of the IR score map has been adjusted to make this distinction more obvious. Right: The H&E image of this sample.

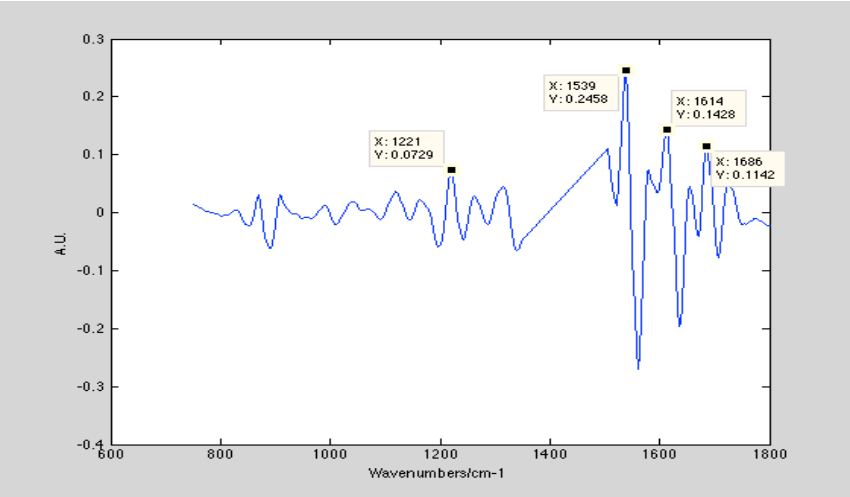


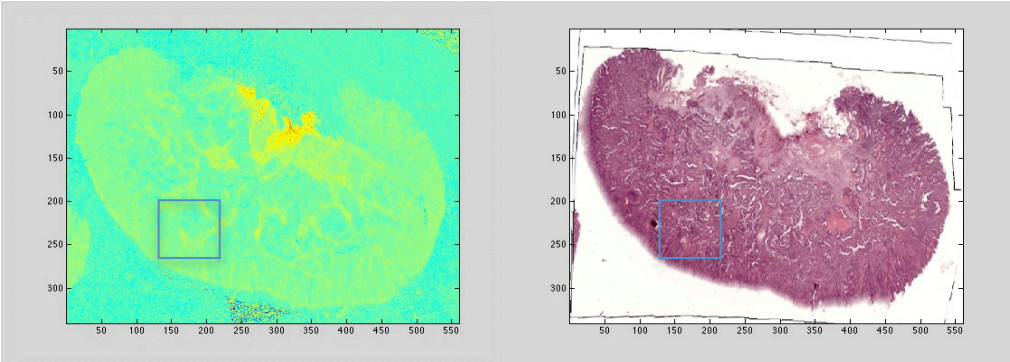

**Figure 4.63.** The 12th PC loading that is evidently related to the submucosa.

From Figure 4.62 it can be seen that epithelial features are identified and this is evident when the H&E for this sample, the right of Figure 4.62, and the IR score image are compared. It was hoped that the application of this image's loadings would enable for epithelial regions within other IR maps to be identified.

On some occasions, this method worked well and an example of good epithelial localization is presented within Figure 4.64. However, it did not work well for all samples, Figure 4.65, and this might be due to the fact that the twelfth component was being used in the identification of the submucosa. As the number of principal components descends, smaller variations in the data are more likely to be identified. Therefore, there is an increased likelihood that noise is being described in higher components. It may also be the case that the statistical relationship identified in Figure 4.62 might be specific for the analysed image.

Further limitations to the imaging of whole tissue sections with a low number of coadded scans is that noise will affect the data. Noise might be present for a multitude of different reasons and the literature suggests that it maybe due to thin tissue samples or because of the rough edges of dense epithelial tissue [101].

It is evident from Figure 4.65 that instrumental and processing defects are causing the noise associated with the images spectra. This is because lines are present within the image and this was again probably due to the fact that only two coadded scans were used. However, the size of the tissue samples prevented longer measurement times and so alternative methodologies were sought.




A B

**Figure 4.64.** The location of a samples epithelium after the application of a targets PCA loads. A) This is the predicted score thought to be related to the difference between the submucosa and everything else within the sample B) The H&E of this sample. The blue box in A&B represents where similar features can be found between images.

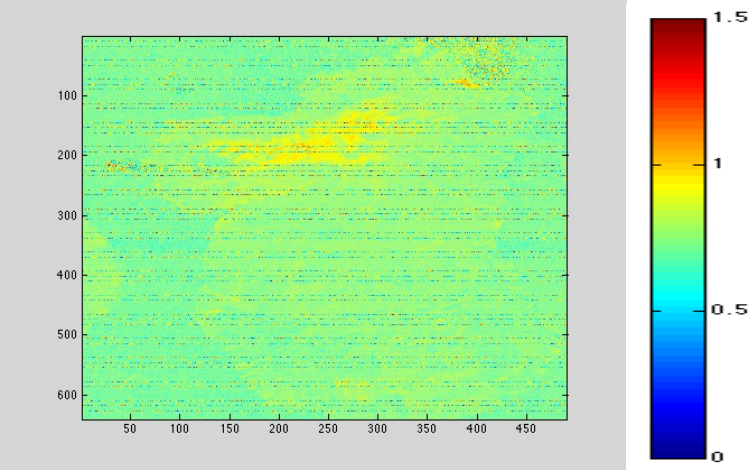



**Figure 4.65.** The poor localization of the epithelium within this sample when the loads of the target were applied.

4.3.2.2 Non linear intermodal registration

As a reliable training loading could not be derived, other methods of representing the epithelium were sought. A novel alternative was to register together images of histological stains and IR chemical maps. In this instance, Haematoxylin and Eosin (H&E) and Cytokeratin (CK) tissues sections were used in the segmentation of epithelial regions.

The tissue used for staining was cut 5µms above and below the section designated for mapping with IR spectroscopy. Therefore, it was thought that because the distance between the different sections was small that they would share similar morphological characteristics. The further registration of these images together means that histological features could be represented by multiple modes. This allowed the CPCA segmentation of the epithelium from the standardly stained images to be used in the collection of epithelial specific spectra. These derived spectra could then be used further within a multivariate statistical model.

To implement this technique the standard histology and IR images were converted into binary masks so that registration could be brought about. This is because it is hard to define points of correspondence between images that are derived from different modalities. Therefore, PCA and K means clustering was used so that a Sum of Squares Difference (SSD) metric between binary images could be optimized.

Optimization was achieved by using a B spline deformation grid and this facilitated accurate registration. A similar registration technique has been used elsewhere within the literature and most recently it has been used with prostate tissue sections, albeit in the derivation of linear transformations [46]. However, it should be noted that the

application of linear registrations with histology images is limited and this is because of the deformations tissue sections undergo when they are being processed[27].

Although linear transformations are not appropriate, they should not be neglected. This is because they enable for further more accurate non-linear registrations to be implemented [26], [149]. In this instance, Principal Axis Transformations (PAT) facilitated linear registrations and these enabled for the accuracy of the non-rigid image deformations to be increased [38].

The accuracy that can be achieved using the implemented algorithm is demonstrated in Figure 4.66. In this image it can be seen that the parallel blue lines between a histological feature, present within both the IR and histology image, are perfectly horizontal. Therefore, indicating that a good registration has taken place. In this instance, the deformations were applied to the histology images as this increases computational efficiency.

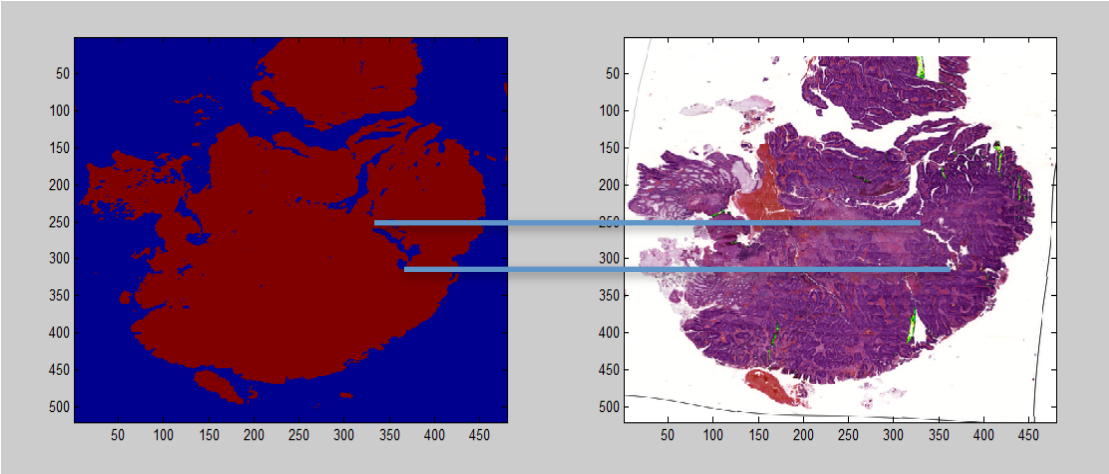



**Figure 4.66.** The results of the non linear registration implemented by this research. On the left is a binary representation of an IR map. On the right is the H&E section that has been cut consecutive to the IR map. The horizontal blue line represents the accurate overlap of features between images that have been transformed.

However, this methodology had several limitations that related to the loss of tissue samples during histological processing. For example, the novel segmentation technique that this research developed required there to be both an H&E and CK image for every tissue sample. Therefore, if an image of an H&E or CK tissue section was missing, other more unreliable segmentation methods had to be implemented. In these cases, PCA was performed on the staining image that was available for a tissue sample and the identified epithelial regions were used in the collection of epithelial specific spectra.

In worst case scenarios, both corresponding H&E and CK images for an IR map were lost during histological processing. This meant that images of stained tissue sections further away from the IR image had to be used. This caused the fidelity of overlapping features within images to suffer. This is evident from Figure 4.67. This is an H&E section that was four levels, representative of a distance of 20µm, below the IR image that was acquired for this sample. It can be noted from this that there are significant compression artifacts and this prevented the reliable collection of IR spectra.

Therefore, further investigations into this methodology were not carried out since epithelial specific spectra could not be reliably sampled from within all registered images.

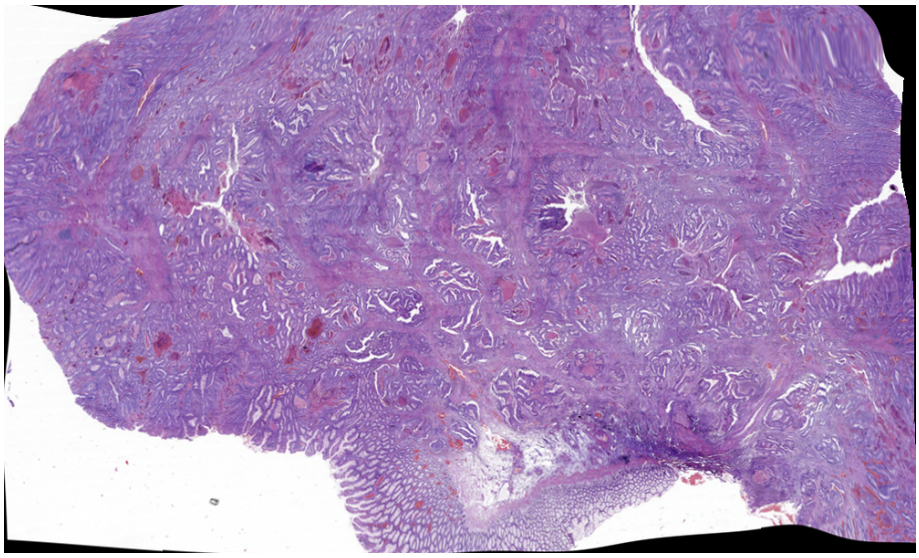


**Figure 4.67.** The unreliable deformations brought about by using sections that were not consecutive to the one mapped with IR.

4.3.2.3 Glandular localisation within unstained tissue sections

Complications with the analysis presented so far include the presence of noise within images and the unreliable overlap of intermodally registered image features. Therefore, as a way of circumventing these problems, a senior Pathologist identified epithelial regions within images of stained tissue samples. The further transformation of these images with respect to low-resolution IR maps enabled for epithelial regions to be located and sampled using a high number of coadded scans. The regions as identified by a pathologist are presented within Appendix C. In this investigation, the time taken to acquire these images meant that only one of these regions could be measured with IR.

In the first instance, IR images were acquired using a spectral resolution of 64cm$^{-1}$ wavenumbers and a spatial resolution of 50µm. Further to this, the map's corresponding image of a stained tissue section was registered with respect to it. This enabled for the coordinates of the epithelial pixels in the H&E or CK images to be used in the acquisition of high resolution spectral maps. As smaller image regions were being acquired in this instance, a higher number of coadded scans (124) were used. In this study, images were acquired using a 25µm spatial resolution and a spectral resolution of 4cm$^{-1}$ wavenumbers.

Since only approximate epithelial locations were required, the exact overlap of features was not necessary. Therefore, circumventing some of the previous problems. It should also be noted that the tissue section analysed for each sample by IR was different to those that were used previously. This allowed for the quality of the histological processing to be evaluated.

From these maps it was determined that tissue sample processing was having an effect on the quality of the recorded data. This is evident from Figure 4.68 where it can be seen that within a samples first six principal components, no morphological structures were being described. It was also evident for this case that the signal to noise ratio of the acquired spectra was poor, even though a high number of coadded scans had been used, Figure 4.69.

In this investigation, as noted, the tissue sections analysed were different to those that had been measured previously. The samples that produced poor quality data were the same for all studies and so it is likely that histological processing introduced uncharacterised defects into some of the tissue samples.

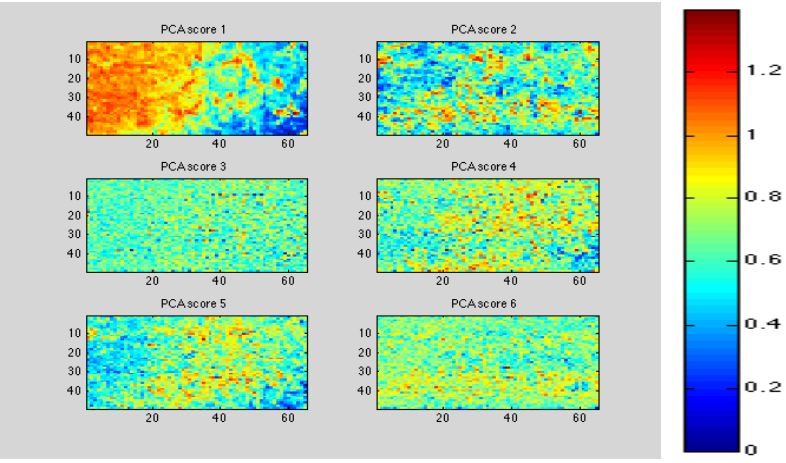



**Figure 4.68.** The lack of morphological structure within the first six PCA scores for an IR map of a gland within a cancer section.

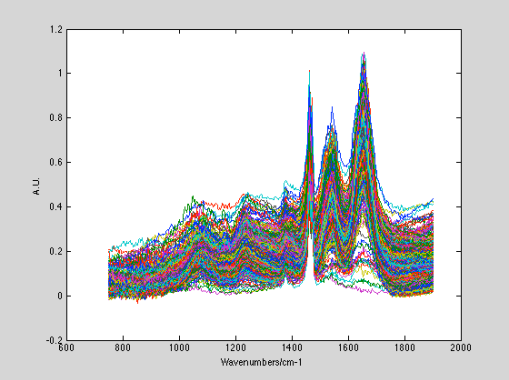



**Figure 4.69.** The problem of noisy spectra within IR maps taken for individual glands within tissue sections.

4.3.2.4 Digital staining

It is obvious from the analysis presented so far that neither the IR data nor standard histology images could be used directly in the derivation of epithelial specific spectra. However, by combining the two modalities a new method of epithelial localisation was established. The developed technique is presented within Figure 3.6. This new methodology uses simple linear regression to correlate the PCA scores of IR images with those of images of standardly stained tissue samples. The technique is very accurate and this is evident from the digitally stained image presented within Figure 3.3.1.1.1. The IR images used in this section are the same as those that were used within section 4.3.2.2.

In this instance, a regression model was built between images where the registration could be visually seen to be very accurate. The regression coefficients from this model were then used to produce functions that could be used in the predictive staining of other tissue samples.

The feature the functions were identifying can be understood by visualising the histology images PCA scores that were used in the regression model. In this instance, two regression models were made. One set of functions allowed IR images to be digitally stained with H&E, Figure 4.70 & 2. The other set of functions facilitated the prediction of E-Cadherin staining, Figure 4.73.

The function used for H&E staining is presented within Figure 4.70. The wild variations within the first loading can be related to the random oscillations of the background slide and the weak signals of paraffin that are accentuated when normalized.

The peaks of the other loadings cannot be observed from Figure 4.70 and so are separately identified in Figure 4.71. It can be seen that the negative peaks within the loadings presented on the left of Figure 4.71 relate to the haematoxylin stain and the

positive peaks relate to everything else. This is because the haematoxylin staining colour has negative values in the histology PCA score that was associated with this function, left of Figure 4.72.

Therefore, the negative peaks of this function relates to regions of the IR spectrum that are associated to the negative values of this score. For similar reasons, the negative peaks within Figure 4.71 correspond to the eosin stain and this is evident when the right of Figure 4.72 is observed.

Within the left of Figure 4.71 it can be seen that dominant peaks for the haematoxylin stain can be found at 1740, 1671, 1545, 1194, 1146, 1092, 990 and 939cm$^{-1}$ wavenumbers. A search within the literature would suggest that the peak present at 1740cm$^{-1}$ wavenumbers can be assigned to the C=O bond stretching of the membranous phospholipids [94], [168], [172], [174].

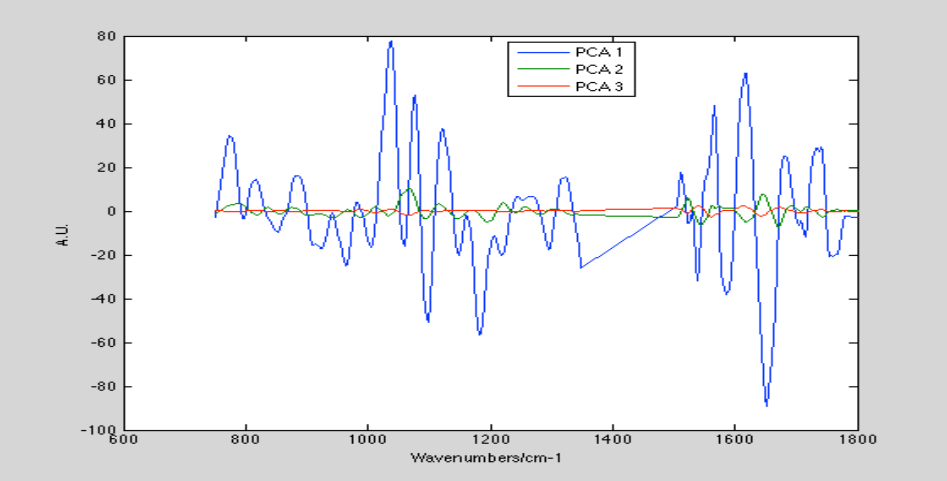



**Figure 4.70**. The function that enables the H&E staining of the IR images to be predicted.

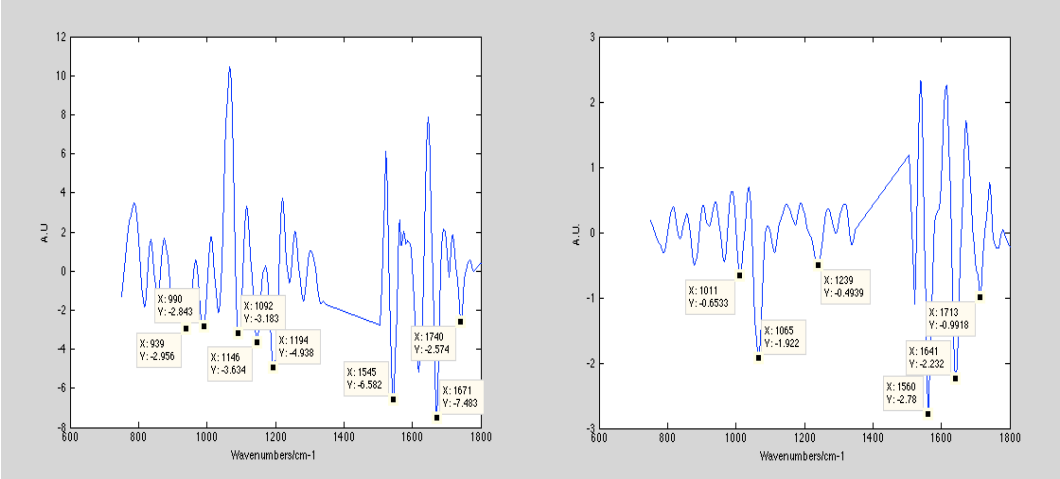



**Figure 4.71.** The predicted stain loadings that enabled the haemotoxylin (left) and eosin (right) stains of H&E to be predicted.

The peak at 1545cm$^{-1}$ wavenumbers can be related to the amide II band of proteins [175]. The band at 1146cm$^{-1}$ wavenumbers maybe related to membranous carbohydrates [139] and the band at 1092cm$^{-1}$ is probably related to nucleotide molecules [170]. The peaks present at 990, 939 and 1671cm$^{-1}$ wavenumbers could not be assigned. However, a similar peak to 990cm$^{-1}$ wavenumbers has been noted within the literature as being associated with colonic cell lines[176].

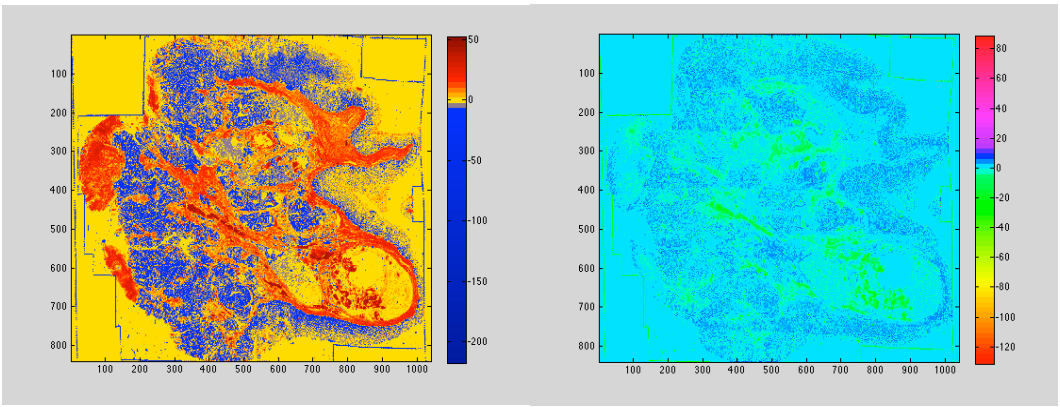


**Figure 4.72.** The PCA scores of the histology images that were regressed against those of the IR image. It should be noted that the colour values have been adjusted so the distinction between features is made clearer.

For eosin the dominant peaks were found at 1713, 1641, 1560, 1239, 1065 and 1011cm$^{-1}$ wavenumbers. The peak at 1641cm$^{-1}$ wavenumbers can be related to the amide I peak of proteins [175], the peak at 1239cm$^{-1}$ wavenumbers was thought to be related to the amide III band of proteins within connective tissue[92]. The bands at

1065 and 1011cm$^{-1}$ wavenumbers could not be assigned but maybe related to carbohydrate molecules [139].

Confidence could be asserted that the model was accurately describing the biochemistry associated with particular stains because it performed consistently across all samples, Appendix D and Electronic Appendix B. Further proof also comes from the fact that the peaks contained within these functions are highly related to the histological components that the H&E stains identify.

A similar regression model was made between an IR image and an image of a tissue section that had been stained with the E-Cadherin antibody. The peaks associated with the stain are presented within Figure 4.73. From this it is possible to ascertain similar peaks to the function that enabled digital haematoxylin staining, left of Figure 4.71. For instance the phospholipid peak is present at 1740cm$^{-1}$ wavenumbers [94] and a similar peak around 1671cm$^{-1}$ (1674cm$^{-1}$) wavenumbers was also present. There were also some new peaks present that might be indicative of E-Cadherin staining features. However, it is unlikely that the E-Cadherin protein was being modeled and this is due to the numerical aperture of the IR system.

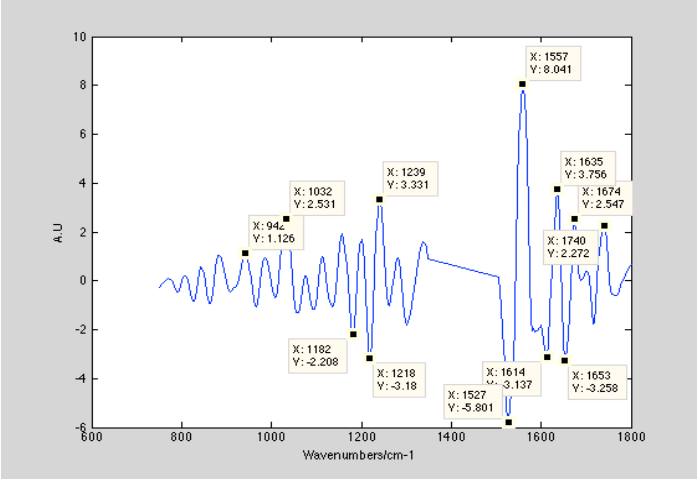



**Figure 4.73.** The predictive stain loadings for E-Cadherin. Those with a positive score relate to the E-cadherin stain and those with a negative score relate to everything else.

The new peaks within the loading for the E-Cadherin stain can be found at 1635, 1557, 1239, 1032 and 942cm$^{-1}$ wavenumbers. The peak at 1635cm$^{-1}$ wavenumbers is probably related to the β sheet structure of proteins [139] and the peak at 1032cm$^{-1}$ wavenumbers may be related to mucin [90].

In this instance, the peak at 1239cm$^{-1}$ wavenumbers seems to be strongly associated with E-Cadherin staining, whereas previously it was associated with eosin. The current

literature provides no alternative explanation for this peak other than it is related to the amide III bands of proteins of connective tissue [92]. It should also be noted that the peak present at 942cm-$^{1}$ wavenumbers could not be assigned.

The E-Cadherin image PCA scores that enabled for the above function to be produced are presented within Figure 4.74. From this it can be seen that only two scores were representing histological features as no informative information is contained within the last score.

However, it should be noted that there were non zero coefficients present for the third component. For this component the most significant regression weights were mainly associated with higher score values and it is likely that these were describing the noise of the data.

Example images of the H&E and E-Cadherin digitally stained IR images are presented within Figure 4.75. All of the digitally stained images for the measured IR maps are presented in Appendix D and it can be noted that in most instances they are highly comparable to those within Appendix C.

However, in a few cases, the predicted staining was not as expected and this is because of noise affecting the images. As previously discussed, the noise present within images maybe due to a multitude of different reasons [101]. Although, it can be noted from Figure 4.76, that in some of these images epithelial regions could still be visualised.

In Figure 4.76, the predicted staining for a sample and its standard histological counterpart are overlaid. The blue boxes in these images represent regions where comparable epithelial regions can be seen. Another example of where the accuracy of this method in segmenting the epithelium from within IR images is presented in Figure 4.77. In Figure 4.77 B, which is the segmented IR image, the epithelial regions identified are highly comparable to the samples CK stained equivalent, Figure 4.77 C. This is even though the epithelium is not immediately obvious from the digitally stained image, Figure 4.3.4.2.8 A.

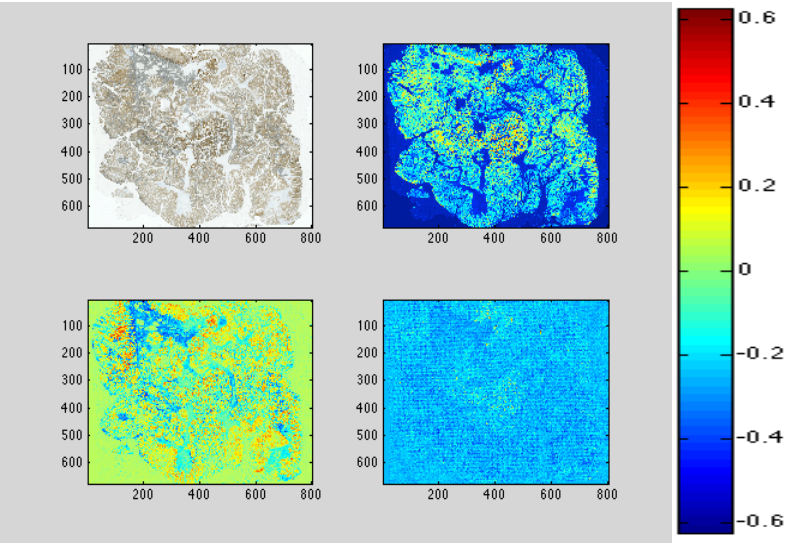



**Figure 4.74.** The PCA scores of an image of an E-Cadherin stained tissue section that was registered to its corresponding IR chemical map. Top left) the original E-Cadherin stained image. Top Right) PCA score one possibly depicting the difference between intense and weak colours. Bottom Left) The scores representing the difference in colour between the connective tissue and the epithelium. Bottom Right) This score was assigned to random noise. Colour bar only corresponds to the PCA scores.

In most samples the predicted staining allowed for the epithelium to be accurately located. However, it did not function for all. This was true for two samples of cancer and because of this their spectra were not included within subsequent multivariate models. The cancer cases that were left out of further analysis are images P & W in Appendix C.

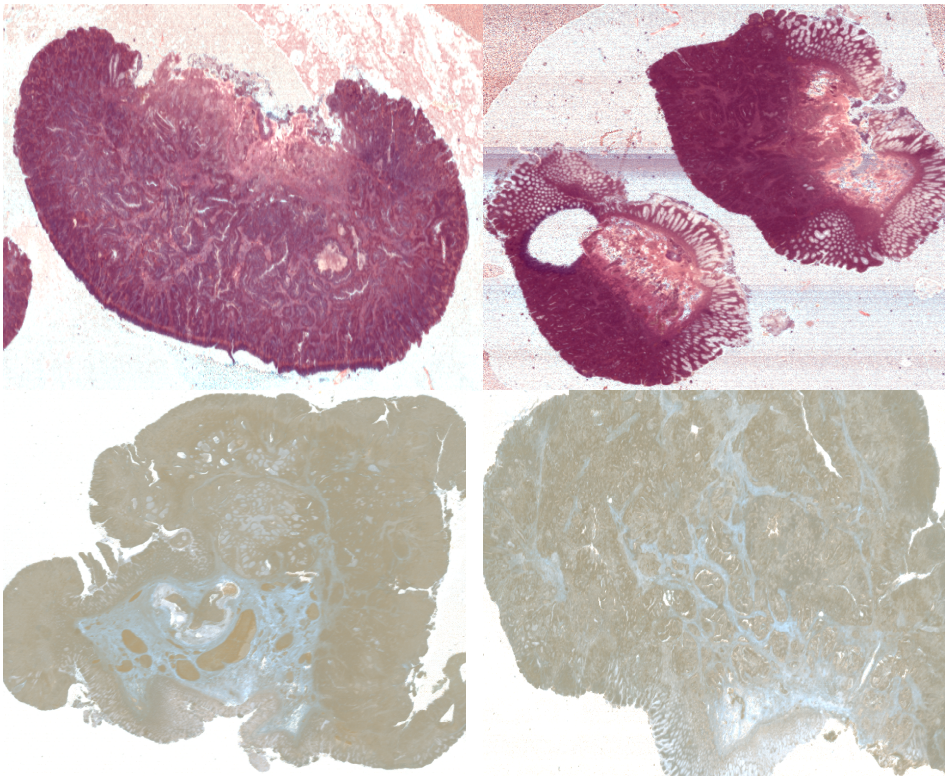


**Figure 4.75.** The predicted H&E and E-Cadherin staining for four IR maps. The corresponding images for these within in Appendix C are, starting from the top in a left direction, images L, V, U and F.

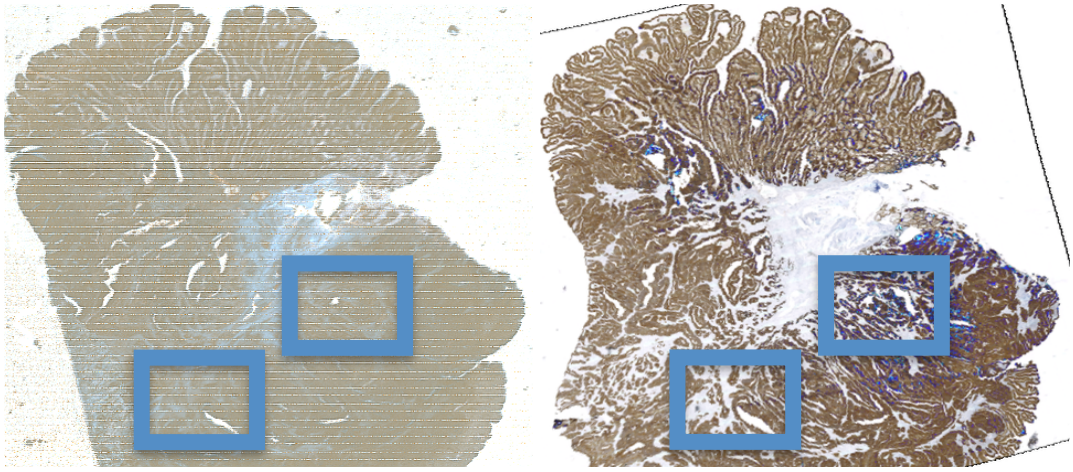
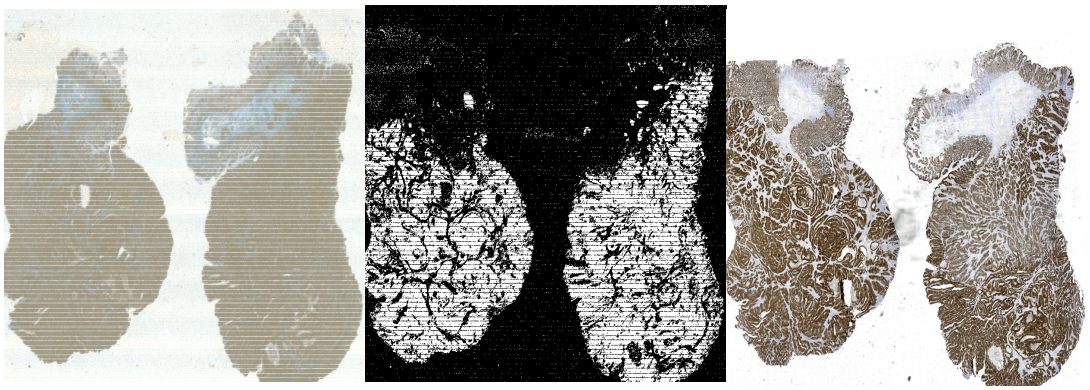


**Figure 4.76.** The predicted E-Cadherin staining for a case of cancer. On the left, the predicted staining is overlain with its corresponding image of a Cytokeratin (CK) stained section. On the right the CK staining image is presented. The blue box highlights regions where the overlap of epithelial regions can be seen.



A B C

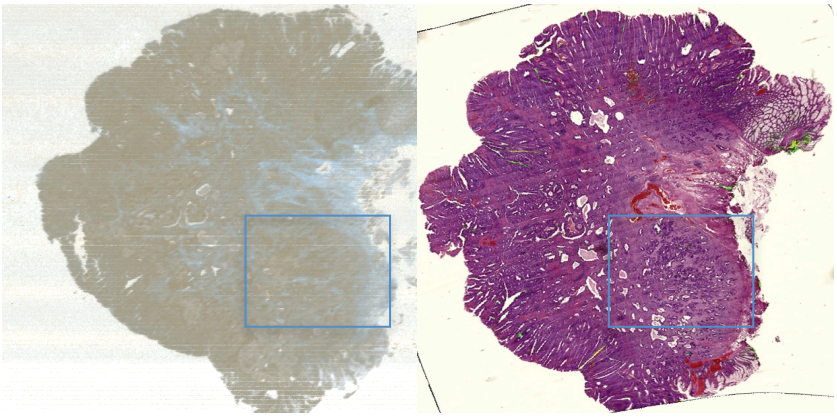
**Figure 4.77.** The accurate demarcation of a tissue sections epithelium by predicting its E-Cadherin staining characteristics. A) The digital E-Cadherin staining for a cancerous sample. B) The manually thresholded PCA scores of the image presented in A. C) Cytokeratin staining for this case of cancer. Comparisons of B with C reveal that the epithelium can be segmented via digital staining.

Spectra could not be derived from these images because no comparable features with their standardly stained equivalents could be found, Figure 4.78. The blue box presented on Figure 4.78 indicates were epithelial tissue might be located but when the two are overlain no precise overlap can be seen. Therefore, no epithelial specific spectra could be collected from this sample. It was thought that the lack of alignment was attributable to human error where the image in B was erroneously marked as being contiguous to A.

4.3.2.5 Multivariate modeling

The images where the predicted staining was successful are presented within Appendix D. From these it can be seen that the best localization of the epithelium is within the digitally stained E-Cadherin images. Therefore, these images were manually thresholded, Electronic Appendix C, so that epithelial spectra could be collected. These were used further within a multivariate statistical model. In this instance, manual threhsolding was appropriate because of issues relating to noise.

Confidence can be asserted that the epithelial regions are being segmented as the further analysis of the segmented regions against their corresponding histology counterparts enabled for accurate correspondences to be made, Appendix E. In these images the segmented images are overlain on their standard histology counterparts. This was so the accuracy of the segmentation was highlighted.



A B

**Figure 4.78.** The ambiguous location for the epithelium within one digitally stained E-Cadherin image of cancer. A) The predicted E-Cadherin staining for an IR map of a cancerous sample. B) The standardly stained counterpart of the image presented in A. The blue box in both of these images indicates were some glandular tissue might reside but no precise overlap could be seen. Therefore, no spectra were collected from this sample for further use within a multivariate model.

From these segmented regions, epithelial specific spectra were collected and used with a PCA fed LDA model. In this instance, regions selected for inclusion within this model were those that a Pathologist had marked as being representative of EM and cancer, Appendix C. The relation of these regions to similar locations on the predicted staining images facilitated the collection of pathology specific spectra.

However, it should be noted that if the predicted staining was thought to be too poor in the identified regions, then no spectra were collected from these areas. This is because the quality of these spectra was likely to be poor. For those regions that were sampled, a total of 154,338 and 70,818 epithelial specific spectra were collected for cases of EM and cancer respectively. The numbers of spectra collected from each sample is presented within Table 4.79. Large differences in sample sizes can be explained by the fact that, in this analysis, the data that was collected was solely from the tissue samples submucosal epithelial islands. In the cancer cases these islands were sometimes small and covered only a few pixels.

The collected spectra were also subject to preprocessing and in this study this amounted to taking the second derivative of the spectra by using a third order 15

point Savitzky-Golay filter; removing the spectras paraffin peaks and finally normalizing and mean centering them.

Subsequent to analysis by PCA the first 20 components were scrutinized by ANOVA. This was so the most discriminatory components could be found. Further to this, LDA was used to minimize variation within groups and maximize the variation between them. In this way, automated LDA loading functions were produced that can be used in the automatic differentiation of disease states.

In this instance, 19 principal components were determined by ANOVA as being significantly different. The high number of discriminatory components identified might be due to the presence of low frequency noise and because a second derivative filter was used in the spectral processing.

It is unlikely that high frequency noise was affecting the analysis because the epithelial specific spectra were subject to quality assurance tests. These tests involved implementing the Q statistic, Figure 4.79, and z scoring. Z scoring refers to removing those spectra that were two standard deviations above or below the integrated amide peak's mean [177].

From the right of Figure 4.79 it can be seen that the raw spectra contained many anomalies. However, when the Q statistics 95% confidence limits, calculated from the first five principal components, were used it can be seen from the left of Figure 4.79 that the noisy spectra were removed.

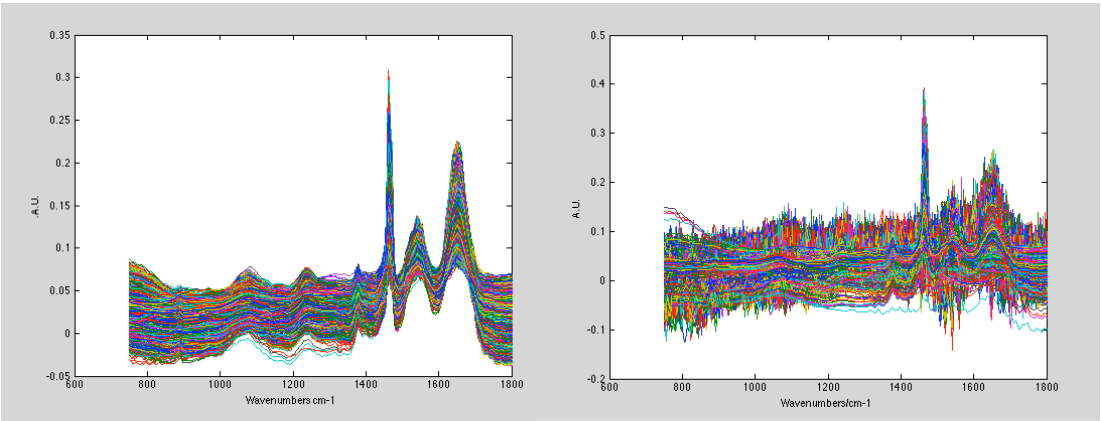



**Figure 4.79.** The spectra for EM and cancer cases that was used further within a multivariate statistical model. Left) the spectra used in the model after quality assurance had been conducted. Right) The raw and evidently noisy spectra that were acquired for tissue sections.

In this instance, the LDA model correctly predicted class membership for the spectra of cancer and EM cases 88% of the time. However, this does not reflect how the model

performed with regards to correctly identifying the pathology of tissue samples. For instance, all of the misclassification could be occurring in one sample.

However, it was found that this was not the case. This was determined from using majority classification in the derivation of a tissue samples pathology. Majority classification refers to the assignment of class membership when more than 50% of a samples spectra were within a given pathology group. When using this classification criterion, the model correctly predicted class membership for EM samples 85% of the time and cancer cases 90% of the time.

The individual classification results are presented within Table 4.79. From this table, it can be seen that the training model is highly sensitive with regards to identifying cancer. However, for the EM cases it can be seen that the correct classification of the spectra per sample is highly variable. This could be due to the spectroscopic differences between EM and cancer being very subtle.

This would have meant that the low number of coadded scans used might not have resolved all discriminatory peaks. Therefore, in future research a higher number of scans for each of the IR images pixels should be used. However, it should be noted that this would cause the sample processing time to be considerably extended.

For those spectra, within Table 4.79, where classification correlated with pathological assessment, the discriminatory biochemistry can be determined. The LDA loading that describes the differing biochemistry between cases of EM and cancer is presented within Figure 4.80.

From the LDA load that is presented within Figure 4.80 it can be seen that for the EM cases there were two very significant peaks. The green numbers that are present on the figure identify these. The first peak, green number one, is present at 1746cm$^{-1}$ wavenumbers and the second peak, green number two, is present at 1062cm$^{-1}$ wavenumbers.

In this instance, the peak present at 1746cm$^{-1}$ wavenumbers has been assigned to the C=O vibrational modes of the phospholipids contained within the membranes of the epithelium [94], [170]. The other peak present at 1062cm$^{-1}$ wavenumbers was assigned to the C-O stretching mode of DNA [89], [170].

The presence of these peaks in the identification of adenomatous tissue, which some cases of EM are an example of, is in keeping with the current literature. For example, it has been determined that there is a difference in the intensity of phospholipid peaks between colonic pathologies [84], [85], [94]. Although, it should be noted, that on

occasion, some of the peaks identified in the literature are on a different basis to what is presented here.

For the cancer cases, the significant peaks identified were those found at 1719, 1698, 1086 and 1029cm-$^{1}$ wavenumbers. The peaks present at 1719 and 1698cm$^{-1}$ wavenumbers, indicated by the red number one and two on Figure 4.80, were hard to identify. However, a search within the literature seems to suggest that they may be related to nucleotide molecules [178].

The red three on Figure 4.80 represents a peak which is present at 1086cm$^{-1}$ wavenumbers and this has been noted within the literature as being due to the vibrational modes of cancerous phosphate groups [176]. Finally, the red number four indicates a significant peak at 1030cm$^{-1}$ wavenumbers and in this instance this was assigned to carbohydrate metabolites [85], [88], [94]. This might be present due to the increased metabolic turnover of cancer cells but further investigations are required [95].

There are a host of other peaks within the presented LDA loading but they are not as significant as the peaks identified above. Therefore, they were not analysed further because of the limited impact they were having on the final analysis. Caution should be exercised when reviewing the analysis above as the LDA loading presented is specific to a model where all of the recorded dataset was used. Therefore, it might be biased towards it.

As an alternative to the above analysis, significant peaks for each pathology could be found from within the models PCA loads. However, given that second derivative spectroscopy was used and because a large number of significant components were identified this would be difficult. Therefore, analysis of this sort was not conducted.

To accurately determine the significance of the peaks identified in Figure 4.80 with respect to discrimination, more thorough investigations are required. For instance, in future research the discriminatory peaks identified upon cross validation could be quantified. Although consensus peaks within cross-validated models were not identified, the procedure was still carried out. Cross-validation refers to the testing of the generalisability of the multivariate model presented above with respect to new samples.

**Table 4.2.** A table displaying the number of spectra collected for each sample and the training and cross validation percentages for the PCA fed LDA model.

| Pathology | Number of spectra | Training set classification/ % | Cross validated classification/ % | Pathology | Number of spectra | Training set classification/ % | Cross validated classification/ % |
|---|---|---|---|---|---|---|---|
| **EM** | 6011 | 96 | 94 | Cancer | 4231 | 100 | 99 |
| **EM** | 7475 | 94 | 91 | Cancer | 4619 | 99 | 95 |
| **EM** | 6128 | 85 | 80 | Cancer | 7860 | 98 | 95 |
| **EM** | 1967 | 97 | 97 | Cancer | 12250 | 96 | 86 |
| **EM** | 393 | 99 | 99 | Cancer | 1289 | 100 | 99 |
| **EM** | 421 | 44 | 44 | Cancer | 343 | 97 | 97 |
| **EM** | 371 | 26 | 25 | Cancer | 574 | 97 | 96 |
| **EM** | 1419 | 83 | 81 | Cancer | 1877 | 83 | 78 |
| **EM** | 2396 | 56 | 50 | Cancer | 1814 | 78 | 74 |
| **EM** | 12210 | 84 | 62 | Cancer | 4717 | 55 | 34 |
| **EM** | 5713 | 73 | 57 | Cancer | 2624 | 56 | 45 |
| **EM** | 1378 | 68 | 67 | Cancer | 1702 | 78 | 63 |
| **EM** | 1765 | 87 | 85 | Cancer | 3190 | 99 | 99 |
| **EM** | 25098 | 99 | 98 | Cancer | 1065 | 1 | 1 |
| **EM** | 17549 | 91 | 86 | Cancer | 2307 | 18 | 10 |
| **EM** | 26362 | 91 | 76 | Cancer | 4492 | 97 | 94 |
| **EM** | 4267 | 95 | 95 | Cancer | 6454 | 99 | 98 |
| **EM** | 921 | 37 | 37 | Cancer | 2587 | 97 | 96 |
| **EM** | 2842 | 76 | 71 | Cancer | 3347 | 100 | 99 |
| **EM** | 19924 | 91 | 58 | Cancer | 3476 | 99 | 79 |
| **EM** | 1900 | 65 | 59 | | | | |
| **EM** | 2477 | 48 | 45 | | | | |
| **EM** | 1809 | 94 | 95 | | | | |
| **EM** | 2335 | 100 | 100 | | | | |
| **EM** | 931 | 99 | 99 | | | | |

In this instance, Leave One Region Out Cross Validation (LOROCV) was utilized. This simply involved iteratively leaving the spectra associated with regions that a Pathologist had previously identified, Appendix C, out of the multivariate model. The remaining spectra within the PCA fed LDA model were then used to predict the class membership for those spectra that were left out. Within each iteration of the cross validated model, ANOVA was used in the selection of the most discriminatory components.

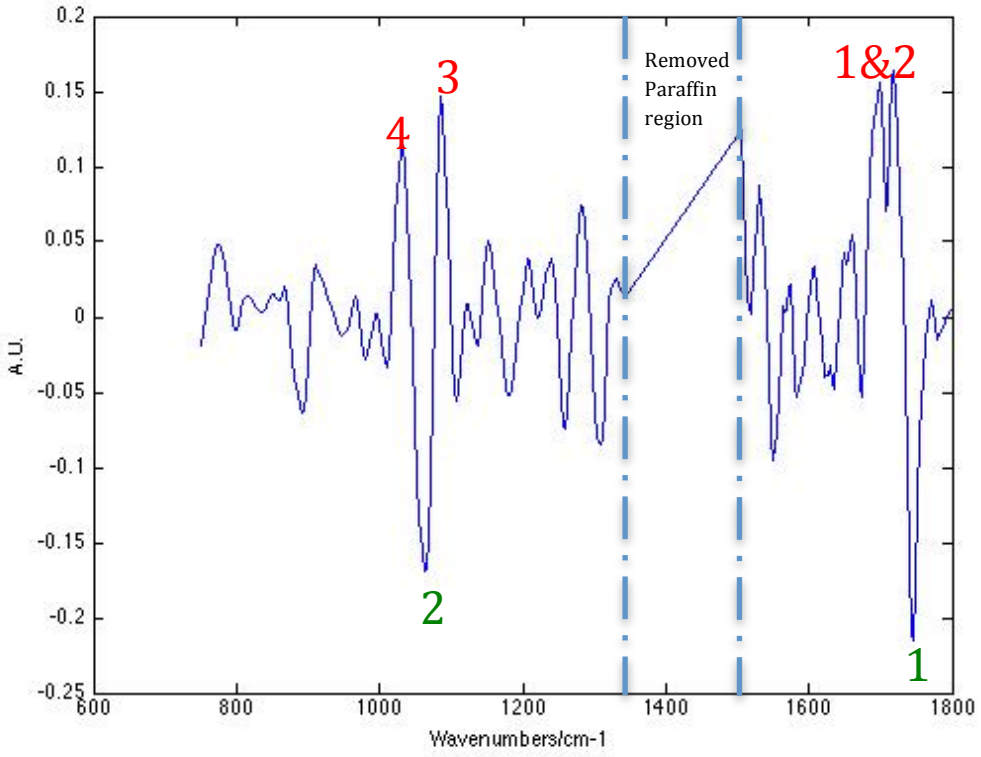



**Figure 4.80.** The LDA discriminatory loads, where the significant peaks are identified by green and red numbers for cases of EM and cancer respectively.

When using the same pre-processing parameters as before and the majority vote classification rules the cross validated samples were correctly predicted 80% of the time in both cases of cancer and EM. This is comparable to the current literature [84], [94], [95] but the analysis presented here has been done so on much larger spectral sample sizes. The cross validated percentages are presented within Table 4.79. It can be seen from this table that for some samples the classification values remain high. However, for other samples there is a drop but this should be expected when cross validation is carried out.

### 4.3.3 Summary

The above analysis has demonstrated that IR spectroscopy has many benefits but great care has to be exercised when it is used in the disease classification process. For instance, this investigation chose to use IR reflective slides with a low number of coadded scans. It may have been more preferable to use the transmission mode of IR spectroscopy as this might have negated some scattering defects [100], [101].

Although, the parameters for this analysis might not have been optimal, this study does come some way in determining the clinical usefulness of probing histological

components with IR. For example, the lack of mucin associated with the adenomatous polyps and some of the cancer cases, Appendix A, means that it is not a good target.

This study also found that it was difficult to accurately differentiate disease states using this compound and this is on the contrary to the current literature [88]. It was the lack of success with mucin that made this study investigate the spectroscopic properties of the epithelium. This follows on from a similar ideology presented within the literature [51] but the final technique presented overcomes the laborious nature of the publicised method.

In this instance, multiple methodologies were tested so that epithelial specific spectra could be abstracted but only one was found to be accurate. In this methodology a novel digital staining technique was developed, Figure 3.7, which allowed the staining characteristics of IR maps to be predicted, Appendix D.

From these epithelial specific spectra were collected and enabled for the cancer and EM cases to be accurately classified 80% of the time upon cross validation, Table 4.1. This is highly comparable to the current literature [84], [85], [94], [95] but it has been done so using specific histological components and has used a much larger spectral sample size. Therefore, the analysis presented might be more meaningful. This is because the inclusion of other histological spectra into multivariate models, which may have been possible in other studies, will mean that is unlikely that biochemical differences between the pathology groups were being described.

A discussion of the all investigated techniques is provided next and this is followed by a conclusions section.

# 5.0 Discussion

The above results demonstrate that both Immunohistochemistry (IHC) antibody staining and Infrared (IR) spectroscopy have the potential to differentiate Epithelial Misplacement (EM) from cancer. In this instance, measuring the continuity of three dimensional (3D) volumes did not facilitate discrimination but it allowed other methodologies to be developed that had application within other studies.

## 5.1 3D volume rendering

It was thought that analysing the continuity of the submucosal glandular tissue within cases of EM would facilitate its classification. This is because the literature postulates that the displaced tissue within the EM cases will remain in connection with their initial misplacement sites [9]. This is on the contrary to cancer, where it is believed that the submucosal epithelial tissue has formed because of metastasis. Therefore, in the cancer cases, the glandular tissue within the submucosa should be disconnected from the samples epithelial surface.

To analyse this potential difference this investigation used image stitching [123], registration [24], [25], [29], [40], [41] and segmentation algorithms [59], [63], [65], [67] to record the number of connected components contained within epithelial specific 3D volumes. If the cancer cases were more disconnected than the EM samples then the number of connected components within these volumes would be higher.

### 5.1.1 Image Stitching

Initial investigations began with the implementation of stitching techniques as these provided images that could be used further with registration algorithms [119], [123]. In this instance, panoramas of tissue sections were constructed using a variety of different methodologies. For instance, one mosaicking technique used the Scale Invariant Feature Transform (SIFT) descriptor algorithm [35] and Normalised Correlation (NCC) [119], [123] in the matching of similar features between overlapping images. However, this technique was not fit for purpose as it was noted that the homologous regions of some tissue sections were problematic, Figure 4.3.

Therefore, methodologies such as phase correlation [124] were implemented as an alternative to the above. The results of using such a technique are presented within Figure 4.4 and it can be seen that it is very accurate. However, the computational complexity of this method prevented its further use. This is because it took a long time to stitch images together.

The poor reliability and time ineffectiveness of the above methodologies meant that automated imaging systems were considered. In this research, the Charged Coupled Device (CCD) of the Perkin Elmer Spotlight 400 Spectrometer (Massachusetts, USA) and Aperio T2 scanners (San Diego, USA), which is part of the Leeds Institute of Molecular Medicine (LIMM), were used.

These technologies enabled the accurate acquisition of whole tissue sections and this is evident from Figure 4.5. Images taken by the Aperio T2 scanners could be used directly within registration algorithms but the raw images acquired by the Perkin Elmer device required a further element of processing. This is because they contained significant vignetting anomalies. However, these defects were easily treatable via the application of a reference image [25].

The corrected and raw images that were derived from each stitching method were used further within registration algorithms. This involved the transformation of images by linear [24], [35] and non rigid means [31], [41], [42] and these enabled the continuity of the volumes epithelial structures to be measured.

### 5.1.2 Image registration

Investigations into registration began with techniques such as the block matching algorithm [24], SIFT descriptors [35] and Moving Least Squares (MLS) [31]. The SIFT and block matching algorithms facilitated the definition of correspondence points between images and these were used further with the non linear MLS image registration technique. It is evident from Figure 4.7 and Figure 4.11 that there were several limitations associated with these techniques that prevented their further use.

For example, the points of correspondence identified by the block matching algorithm, Figure 4.7, were to unreliable to be used in the linear and non linear transformation of images. Linear registrations were investigated because it was thought that they would provide accurate starting points on which other methodologies could improve [74].

For instance, rigid transformations can aid B Spline grid algorithms that deform images via the minimization of a Sum of Squares Difference (SSD) metric [26], [29]. When using this algorithm, the optimisation that is performed is a result of finding the optimal spline coefficients that enable the smallest SSD value to be found. This is often subject to problems relating to local minima but the preliminary transformations can overcome this.

As the block matching technique could not be used, the SIFT algorithm was used as an alternative in both the linear and non linear transformation of images. It can be noted

from Figure 4.9 that it enabled images to be accurately rigidly aligned. In this figure, linear registration was achieved by minimizing the Euclidean distance between SIFT points that had been refined by the RANdom SAmple Consensus (RANSAC) algorithm [37].

RANSAC uses a parameterised model to estimate linear coefficients that would rigidly overlay points of correspondence with respect to one another. However, the linear model used within RANSAC meant that points relating to similar features between images that had been deformed in a non rigid fashion, due to histological processing, were removed. Therefore, the retained SIFT points after RANSAC could not be used in the accurate non linear transformation of images.

To circumvent this problem, the SIFT points were clustered into four groups and then RANSAC was performed on each one individually. The retained points were then used with MLS in order for the original pose of the tissue sample to be regenerated. A similar idea has been used else where within the literature [36] but in this instance it was found that many erroneous points of correspondence were retained. It is evident from Figure 4.11 that they affected the final registration and this means that more sophisticated refinement algorithms will have to be considered in the future.

Further problems associated with this technique relate to the assumptions used during its operation. In some instances, only a few sparse SIFT points were found for particular regions and this is because the images were resized before they were processed. The resizing of images was necessary because of the computational constraints that exist when working with images in Matlab (Natwick, USA) on a standard desktop PC.

It was thought that the points found in the resized images would represent the deformations that occurred to all features. However, the complex nature of histological processing meant that this was unlikely to be the case.

The technical difficulties associated with the above methods resulted in collaborating with the Leeds Institute of Molecular Medicine (LIMM). This group has developed an accurate image registration algorithm that has a proven track record within the current literature [41], [42].

Their algorithm works by initially rigidly transforming images and then using these with the phase correlation metric in the definition of points of correspondence [42]. A B Spline grid was then used to minimize the distance between these points so that images were deformed with respect to one another. It is evident from Figure 4.12 that

their technique is very accurate and these allowed histological features to be represented in 3D.

However, the epithelial features of interest are not explicitly represented within these volumes and so segmentation algorithms were implemented. In this instance, many different types were investigated. This included segmentation by colour deconvolution [59], [60], spatially corrected Fuzzy C Means (FCM) clustering and active contours without edges [65].

### 5.1.3 Image Segmentation

In all of the segmentation methodologies investigated, it was noted that the non specific staining for the epithelium within Haematoxylin and Eosin (H&E) images prevented this features accurate segmentation, Figure 4.13. The use of these techniques with images of tissue sections that were stained with the Cytokeratin (CK) antibody was also problematic. In these images, the application of the aforementioned techniques preserved undesired features, such as blood vessels (Figure 4.24).

Therefore, a novel methodology was developed that enabled epithelial structures to be accurately represented within 3D volumes. This involved the intermodal registration of contiguous images of tissue sections that had been alternatively stained with H&E and the CK antibody.

The alignment of these images with respect to one another enabled for epithelial specific staining relationships to be identified. Only in the epithelial regions of the fused images were pixels coloured brown and purple due to the CK antibody and H&E staining respectively.

The Consensus Principal Component Analysis (CPCA) algorithm [137] identified this trend and along with the 90% confidence limits of the D Statistic [138], calculated from a training set of epithelial specific super scores, facilitated segmentation. This can be seen from Figure 4.23 and Electronic Appendix A and these show that this methodology was very accurate. Within Electronic Appendix A, movie representations of all 13 cases of EM and 9 cases of cancer that were accurately segmented using this technique are presented.

For this segmentation to work there was the requirement that for each sample there was strong epithelial staining within both its H&E and CK images. This was problematic for one case of cancer, Figure 4.25, and meant that for this sample a 3D volume could not be made. Therefore, no connected components were recorded for this case.

However, this type of segmentation worked for the 2,800 images that make up the volumes found within Electronic Appendix A. The number of connected components contained within these was recorded and used further with the 99% confidence limits of the Mann Whitney U Statistic. This tested the null hypothesis that cancer cases were as connected as EM samples. However, no significant differences were found and so the null hypothesis could not be rejected.

This result would suggest that the cancerous cases were as connected as the EM samples but it might have been the case that the adenocarcinomas analysed here were very early stage. Therefore, complete disassociation from the surface might not yet have been brought about.

The similar number of connected components recorded might have also been due to the experimental design used. In this instance, an automatic image pairing algorithm was used so that corresponding CK and H&E images from a sample could be found, Figure 4.19.

For every sample, a CK and H&E volume was created independently as it is difficult to define points of correspondence between images of these stains. This caused the order of the alternatively stained tissue sections to be lost. Therefore, multiple ways of correcting for this, such as by manual review and record keeping, were implemented but most were found to be to error prone.

This resulted in the development of a technique that allowed for image correspondences to be approximated. This involved selecting and deforming images from a samples CK volume with respect to those derived from its H&E equivalent. In this instance, edge filtered images [28] were used with a B Spline grid [29] in the minimization of a Sum of Squares Difference (SSD) metric [28], Figure 4.19.

The minimum from this metric was used to determine the images that were corresponding. These were then used further with more refined registration parameters before application of the CPCA segmentation technique. This pairing had advantages such as it allowed for correspondences to be found for images of tissue sections where their alternatively stained counterpart had been destroyed. It also negated problems related to human error.

Although, in some instances the image pairs found did not align well upon further registration. This maybe due to tears within the imaged tissue sections and this would have prevented their true correspondences from being found. Those images that did not align well were removed from the dataset and this could have introduced artificial

discontinuities into the segmented volumes. Therefore, more sophisticated comparison metrics such as Normalised Mutual Information (NMI) should be considered in the future when finding image pairs [30].

It should be noted that only a small number of sequential images were removed from a samples image stack, much smaller than the tissue that was lost due to processing [27] but this would still impact the number of recorded connected components. Therefore, the results presented within Table 4.1 should be treated with caution.

Other experimental design problems can be related to the fact that only one out of the many tissue blocks available was volume rendered. It could be the case that a block was chosen for the EM cases where the submucosal glandular tissue had already disassociated from the surface. Therefore, by using more of the tissue block in the volume rendering process, a better representation of the continuity of epithelial structures will be brought about.

However, the usefulness of defining such a metric is questionable. This is evident from the movies contained within Electronic Appendix A. These show that the submucosal epithelial glands in the cancerous tissue are as connected to the surface mucosa as the EM samples.

Although, measuring the continuity did not allow for differentiation the novel segmentation technique developed has application within other studies. Its use is preferable to those available in the literature and this is because epithelial glands are often missed when these techniques are used [65], [67], [70], [71]. It was also less technically demanding as it does not require the use of additional technologies, such as hyper spectral scanners [116–118].

## 5.2 Immunohistochemistry (IHC)

The lack of differentiation caused by measuring the EM and cancer samples epithelial continuity in 3D meant that alternative technologies were considered. Within the current literature, staining with particular antibodies has been found to facilitate differentiation [9], [11–13] .

### 5.2.1 Comparisons with the current literature

Antibodies for E-Cadherin, p53, Ki67, Matrix Metalloproteinase 1 (MMP 1) and Collagen IV have been identified as being diagnostically useful and these were further evaluated in this study. In the case of E-Cadherin, p53, MMP1 and Ki67 comparisons in staining intensity were made between the tissue samples submucosal and surface

epithelial tissue. In the case of Collagen IV, differences in the continuity of submucosal and surface epithelial basement membrane staining were noted.

For this investigation, 10 samples of EM and 10 cases of cancer were used with most antibodies, except when Collagen IV was analysed. For Collagen IV, nine samples of EM were compared against 10 adenocarcinomas and this is because one of the tissue sections was destroyed in transit.

The results produced were compared against those of the literature and were analysed by the 99% confidence limits of the two-tailed Mann Whitney U statistic. In this research, E-Cadherin, MMP 1, Collagen IV and Ki67 all produced statistically significant results but some of these were different to those of the literature.

For example, Yantiss et al found that within 65% of the cancerous pathologies they investigated that there was more intense surface E-Cadherin antibody staining [13]. On the contrary to this, this investigation found that within 50% of the cancerous samples there was a decrease in staining for this region. Therefore, although staining with this antibody was found to be significant, caution should be exercised when interpreting the results it produced.

Another discrepancy that was noted between this research and the current literature was when the results of MMP 1 staining was analysed. In the literature, the MMP 1 antibody has been noted as being capable of staining both stromal, the collective name given to connective tissue, and epithelial cells [13]. In this instance, it was noted that when this antibody was used there was no stromal region staining within the EM cases, Figure 4.33. However, it was noted that there was increased staining for this histological component around 100% of the cancerous cases submucosal glandular tissue.

Given that there was no stromal staining for MMP1 within the EM cases it is not surprising that the Mann Whitney U Statistic found significant differences. The lack of staining in the EM samples could be due to technical defects but this is unlikely as there was positive reaction for this antibody within the epithelium of the EM cases, Figure 4.33. When the pathologies epithelial staining with MMP1 was analysed, no significant differences were found and so the null hypothesis could not be rejected. Therefore, the clinical use of this antibody is questionable.

For Ki67, no significance testing has been carried out previously within the literature and so this made any comparisons difficult. However, the literature has noted that there was increased staining for this antibody within cancerous samples[11]. This

investigation found a similar staining pattern where 50% of the cancerous cases displayed increased submucosal glandular tissue staining. In the EM samples there was decreased submucosal staining for this antibody in 40% of cases. The Mann-Whitney U Statistic found this to be statistically significant but the small sample sizes it correctly classified means that further investigations are required.

The final antibody analysed by this research, which produced significant results, was Collagen IV and the analysis followed the current literature [13]. In this instance, it was noted that there was continuous basement membrane staining for this antibody within 100% of the EM samples submucosal and surface epithelial tissue, Figure 4.35. For the adenocarcinomas, the staining was discontinuous or absent in 100% of cases, Figure 4.36.

Analysis by the 99% confidence limits of the Mann Whitney U Statistic found that this finding was statistically significant. However, this relied upon the subjective assessment of this author and a senior Histopathologist. It is well known within the current literature that there is a large degree of variability between pathological evaluations[96]. Therefore, the analysis of Collagen IV will have to be automated before any solid conclusions can be drawn.

In addition to the univariate testing that was carried out, multivariate statistics, in the form of Hierarchical Cluster Analysis (HCA), was conducted on the observed IHC staining scores. In this instance, it was found that the scores given to Collagen IV and stromal staining with MMP-1 were dominating the observed clusters. Therefore, these were removed from subsequent HCA models so that the discriminatory potential of the other investigated antibodies could be evaluated.

It was found from Figure 4.41 that by using combinations of E-Cadherin, p53, Ki67 and epithelial staining with MMP-1 that 80% of the cancerous pathologies could be distinctly differentiated from their EM counterparts. When using selected combinations of the above biomarkers within new HCA models, only small sample sizes were correctly classified and for most models the observed clustering followed the univariate staining characteristics of the most significant biomarkers, Figure 4.42 & 43.

Therefore, the only model, which did not use the most discriminatory stains within its construction, that brought about robust classification was the one that included the IHC scores given to E-Cadherin, p53, Ki67 and MMP-1 epithelial staining. It can be generally asserted from this model that within the EM samples the submucosal glandular regions displayed decreased staining intensities for MMP-1 and Ki67. When

the other antibodies are taken into consideration, it would seem that for EM there should be no difference in staining between its surface and submucosal epithelial tissue. For the cancer cases, there was a general trend of increased MMP-1 and Ki67 staining with decreased p53 expression. However, it should be noted that this result should be treated with caution because of the small sample sizes involved.

**5.3 Infrared (IR) Spectroscopy**

The use of antibodies with pathological assessment in the determination of classificatory biomarkers is difficult because of the subjectivity associated with current methods of analysis[50], [96]. It was for this reason that more objective techniques were sought to bring about differentiation.

A technique that was thought that would enable this was Infrared spectroscopy as it has been used previously in the accurate discrimination of benign colonic tissue from cancer [84], [85], [88], [94], [95]. In this instance, mucin and epithelial histological targets were probed with IR using novel image processing techniques and the benefits of such an approach are highlighted below.

5.3.1 Mucin

It was thought that mucin would be a good target as it is commonly associated with colonic pathologies. In the current literature, the probing of mucin with IR enabled for the correct classification of nine adenocarcinomas and seven non neoplastic tissue samples [88]. However, the differentiation criteria used by Travo et al is unlikely to be repeatable within a clinical setting. This is because K means clustering was used with IR images in the localisation of a tissue samples mucinous components. They then used the amount each mucin cluster covered in each IR chemical map to classify tissue samples. However, results produced by the K means clustering technique will not be repeatable and the metric they defined will be subject to the size of the mucin pool measured.

Therefore, alternative ways of measuring this feature were devised and this involved measuring 0.3cm by 0.3cm mucin regions from tissue samples with IR. In this investigation, eight regions from four cancer and four non neoplastic tissue samples were measured and the regions chosen for mapping were determined from H&E images. The standard histology sections could be used as they were cut consecutive to the IR maps, Figure 4.48, and their linear registration allowed similar regions to overlap.

Once the mucin images were acquired a novel image segmentation methodology was implemented that facilitated the collection of mucin specific spectra, Figure 4.49, 50 &

52. In this analysis, a separate mucin specific Principal Component Analysis (PCA) loading was used with each pathology. To be truly clinically relevant, the spectra from one image should be used to predict the location of mucin from within all. However, the purpose of this analysis was to judge the benefits of probing mucin with IR and so this methodology was justified.

Initial investigations revealed that paraffin peaks dominated multivariate analysis and so several methodologies were implemented to nullify the spectral contributions of this compound [88], [104], [105], [168]. However, in this analysis, it was determined that digital dewaxing algorithms, such as EMSC [88], [104], added to much volatility to the recorded spectra, Figure 4.55.

Other methods such as chemical dewaxing [105], [168] are not always 100% effective and they may also chemically alter tissue samples [104]. Therefore, the main paraffin peaks were deleted from the recorded spectra as an alternative to those presented above, Figure 4.55.

Paraffin was not the only complication as the recorded IR spectra contained many baseline artefacts. These may have been introduced because this research chose to use IR reflective slides [100], [101]. These require the transflection mode of IR to be used, which is prone to scattering problems. In the literature it has been noted that this causes shifts in peak positions around the amide I region of the IR spectrum but it is thought EMSC can correct for this [100], [101]. However, as noted, EMSC added to much variation into the dataset.

Therefore, baseline artefacts were accounted for by using a Savitzky-Golay 15 point third order second derivative filter. This removed the spectra's baselines but it would not correct for peak shifts [100]. Therefore, any discrimination due to the amide I region of the IR spectrum was treated with caution in this analysis.

The processed spectra were used further within a PCA fed Linear Discriminant Analysis (LDA) model but no biochemical differentiation was achieved, Figure 4.57. In this instance, the discrimination that was being achieved was due to physical defects. It is true that the erroneous regions of the spectrum could have been removed and that the analysis be repeated but this may have biased the model towards the training dataset.

Since no meaningful discrimination could be achieved between the non-neoplastic and cancer cases, it is unlikely that mucin will help aid the classification of adenomas.

However, an investigation of this sort was carried out, whereby mucin specific spectra were collected from four adenomas and four cancers.

In this analysis, xylene dewaxing was used to remove paraffin as the deletion of its peaks from the previous datasets might leave out important biochemical information [102]. Mucin from the sampled areas was identified using a similar procedure as before, Figure 4.58, and second derivative spectroscopy was again used to remove baselines. In this instance, the segmentation was not as effective as it was previously and this was attributable to the acquisition of poor quality spectra, Figure 4.59.

The mucin specific spectra collected were used further within a PCA fed LDA model and the results of this are presented within Figure 4.61. From this figure it can be noted that the PCA loadings with the largest LDA weights correspond to those that contain spectral artefacts. Therefore, the technical difficulties associated with the collection of mucinous spectra means that it is likely to be a poor target for IR.

Other problems include the fact that mucinous pools associated with adenomas do not always survive histological processing, Figure 4.60. This is evident for one adenomatous sample where no mucin pools could be found after processing. However, confidence can be asserted that these were present and this is because they were seen in the standardly stained section of this sample, Appendix B (A). Therefore, the above difficulties demonstrate that the differentiation of EM from cancer by the analysis of mucin will be difficult to implement in a clinical setting.

### 5.3.2 Epithelial specific spectra

Another histological feature that was targeted with IR spectroscopy was the epithelium. This biological entity is abundantly present within colonic tissue samples, as is evident from Appendix A, and therefore provides a consistent target for IR. In this investigation, two studies were conducted where in the first case all of the epithelium within tissue sections was measured. In the second investigation, individual glands were targeted and measured with IR.

Both of these required segmentation algorithms to be implemented, as the retention of spectra related to other histological features will affect multivariate analysis. In the first instance, the spectra contained within the IR images of whole tissue sections were used so that an epithelial specific PCA loading could be found, Figure 4.62. However, no epithelial specific PCA scores were found and this might have been because the epitheliums biochemistry was being explained across many components. The Signal to Noise Ratio (SNR) of some of the images was also poor and this made the reliable identification of glandular tissue more difficult.

Therefore, intermodal registration was used as an alternative in the sourcing of epithelial specific spectra. It is evident from Figure 4.63 that the registration procedure developed in this research was accurate and that it facilitated the reliable overlap of features. In this way, the epithelial glands in the IR data will be spatially aligned with their counterparts in the standardly stained images. Therefore, the CPCA segmentation algorithm, developed in the 3D volume rendering study, could be used in the collection of epithelial specific spectra.

However, there were several limitations associated with this technique. For example, some of the standardly stained tissue sections did not survive the staining process. This meant that the stained sections for some samples were missing. As a consequence of this, images of tissue sections further down the tissue block had to be used.

This affected the assumption that the small distances between sections allowed for histological features to overlap, Figure 4.67. In this figure, it can be seen that the minimisation of the SSD metric between binary images caused unrealistic deformations. Therefore, epithelial specific spectra could not be reliably obtained. Further problems with this technique included partial occlusion. This refers to folds that were present within tissue sections and this caused similar compression problems as seen in Figure 4.67.

It should be noted that multiple low e slides were cut from a tissue block but not all had corresponding H&E and CK tissue sections. This is because much of the block was used for 3D volume rendering and further IHC analysis. Therefore, there was limited tissue available for IR spectroscopy and the use of the samples other low e slides was complicated because of problems relating to the distance between tissue sections.

Although, registration of this sort still facilitated the accurate alignment of some epithelial features between images. For these regions, the pixel coordinates derived from the transformed histology images could be used to derive epithelial specific spectra and this formed the second part of the IR study.

The mapping of individual glands allowed a high number of coadded scans, 124, to be used and it was thought that this would help improve the recorded spectra's SNR. However, the signals contained within these IR images were still of a poor quality, even though a high number of coadded scans had been used, Figure 4.69.

In this new investigation, the tissue sections measured with IR were different to those that were used in the acquirement of images of whole tissue sections. The samples

that contained poor spectra in this methodology were the same as the last. This suggests that these samples were not accurately sectioned. This could be due to thin sectioning [101] but a more in depth investigation is required before any conclusions can be drawn.

The relevancy of studying single glands is also questionable. This is because the selection of individual glands might not be representative of the biochemistry of the entire samples epithelium. This method of analysis is also time consuming as it can take several hours to acquire IR images of individual glands when using a high number of coadded scans. This is especially problematic if the spectra contained within the map are of a poor quality because IR measurements will have to be repeated in other regions.

### 5.3.3 Digital Staining

As a result of the difficulties presented above a novel IR image processing algorithm was developed, Figure 3.7. This involved the non linear registration of standardly stained tissue sections to their IR chemical map counterparts.

The precise alignment of features enabled a simple linear regression to be used in the correlation of the histology and IR images PCA scores. This allowed coefficients to be derived that would relate the IR images variance to that of the standardly stained images. This facilitated the production of functions that would digitally stain images, Figure 4.71 & 73. The benefit of such an approach is that only one pair of IR and standardly stained images, which had been accurately registered, were required.

The model derived from this could then be used to digitally stain other IR images and this negated some of the problems associated with partial occlusion and missing tissue sections. Confidence can be asserted that these functions are representative of the modelled stains because the peaks contained within them are representative of the features that these stains are used to identify.

For example, a peak found at 1740$cm^{-1}$ wavenumbers was found to be associated with the haematoxylin stain and this is representative of membranous phospholipids [92], [94]. The haematoxylin stain can be used to visualise epithelial components and so a peak of this sort should be expected. It is evident from Appendix D, E and Electronic Appendix B that in most instances this allowed the epithelium to be successfully represented.

However, for some images the location of the epithelium was not immediately obvious. Although, upon manually thresholding, it was evident that epithelial regions could be derived from these cases, Figure 4.76 & 77.

In total, 11 cases of EM and 9 cases of cancer were segmented in this way and the binary images made from their digitally stained counterparts are presented within Electronic Appendix C. Although, the presented methodology worked well in the vast majority of cases, it did not function for all, Appendix D E(CK) & L(CK). In Appendix D L(CK) no epithelial regions can be seen, even after thresholding. Therefore, no spectra were derived from this sample for use within a multivariate model. For the image in Appendix D E(CK), epithelial regions could be derived but when comparing this against its standard histological counterpart no accurate overlap of the epithelial features could be seen, the blue box in Figure 4.78.

For the samples were the predicted E-Cadherin staining did work, epithelial specific spectra were collected by manually thresholding these images PCA components. It would have been beneficial to use an automated method but the variable staining within the predicted images prevented this from being achieved. The further acquirement of more refined IR maps will enable automated segmentation methods to be implemented and will allow a technique of this sort to be used within the medical arena.

In this instance, the digital stained H&E images were not used further except for ensuring that no spectra were being collected from blood vessels. Although, in the future a combined H&E and E-Cadherin function should be made as this will be very specific for the epithelium.

From the segmented images, regions identified by a pathologist were selected and the spectra associated with these were collected. This spectra's second derivative was taken, their paraffin peaks were removed, and they were normalised and mean centred. These were then used further within a PCA fed LDA model, where the 19 principal components chosen for analysis were determined by ANOVA [84], [94], [143], [179]. The high number of components found is likely to be due to a combination of low frequency noise amplification [140] and because only two scans were coadded during the acquisition of IR data.

In total, 25 regions were taken from the EM IR maps and this equated to over 150,000 spectra being collected for these cases. For the cancerous samples, 20 regions were analysed and these contributed over 70,000 spectra to the multivariate model.

When all of the spectra were used within the PCA fed LDA model the cancerous and EM regions were correctly classified 85% and 90% of the time respectively. This classification is comparable to the current literature [84], [85], [94] but in this instance it has been achieved using a much larger spectral sample size.

Upon Leave One Region Out Cross Validation (LOROCV), the cancerous and EM samples were correctly classified 80% of the time. The determination of the discriminatory variables within the cross validated model is difficult and requires further investigation. This could potentially be achieved by using techniques such as autocorrelation and CPCA [137].

However, an estimate of the discriminatory biochemistry can be understood from analysing the LDA loads of the model where all the spectra was included, Figure 4.80. From this figure it can be seen that for the cancer cases, the classificatory peaks were found at 1719, 1698, 1086 and 1030cm-$^{1}$ wavenumbers. Some of these peaks were hard to identify [178] but some are comparable to others that have been found in the literature.

For example, the peak at 1086cm$^{-1}$ wavenumbers has been noted within the literature as being due to the vibrational modes of cancerous pathologies phosphorous compounds [176]. The carbohydrate metabolite peak found at 1030cm$^{-1}$ wavenumbers was thought to be representative of the cancers increased metabolic turnover [85], [88], [94]. This is comparable to the literature but further investigations are required [95].

For the EM samples, the peaks that were significant in its classification can be found at 1746 and 1062cm$^{-1}$ wavenumbers. In this instance, the peak found at 1746cm$^{-1}$ wavenumbers was assigned to the C=O vibrational modes of phospholipids [94], [170] and the peak at 1062cm$^{-1}$ wavenumbers was thought to be due to DNA [89], [170].

The association of the phospholipid peak with the EM cases was thought to be important because it indicates that the cancerous pathologies are more fluid in nature. The reduced presence of this band in cancer might be representative of metastasis but further investigations are required. The peak at 1062cm$^{-1}$ wavenumbers might be representative of a nucleotide configuration specific to EM. However, in the literature, peaks in this region have also been related to phospholipids and so may add more credence to the argument surrounding fluidity [139].

### 5.4 Summary

The analysis presented above seems to suggest that further investigations should concentrate on the automated analysis of IHC slides [180] and IR spectroscopy [50], [51], [84], [85], [94], [95]. This is because the investigation into 3D volume rendering found that there were no significant differences between pathologies when their epithelial continuity was analysed. However, this is but only one way in which the segmented volumes can be used. If this modality is pursued in the future then the analysis of the growth characteristics of the submucosal glandular tissue within the EM samples should be undertaken [10].

For IHC staining, many of the results presented were on the contrary to the current literature [11–13], except for Collagen IV. This antibody was found to be useful in the discrimination of EM from cancer but the subjectivity associated with this analysis means that quantitative methods of assessment need to be developed [180].

In other investigations, it was determined that IR has some discriminatory potential. This is because LOROCV correctly classified the cancer and EM cases 80% of the time. This result is highly comparable to the current literature but has been done so using a much larger spectral sample size [84], [85], [94], [95]. Therefore, further studies should be conducted into this modality using more refined resolutions with a higher number of coadded scans.

This section has highlighted the benefits and pitfalls of the developed methodologies and indicates that more work is required. In the next section, the main conclusions of this research will be drawn together and this will be followed by suggestions for future experimentation.

# 6.0 Conclusions and Future work

The above results and discussion demonstrate the potential of many adjunct technologies. The main findings of these investigations are presented below and finally suggestions for future work are discussed.

### 6.1 3D Volume Rendering

The literature postulates that measuring the continuity of epithelial structures within cases of Epithelial Misplacement (EM) will bring about their differentiation from cancer [9]. However, this claim was found to be unsubstantiated. In this instance, segmentation algorithms were used to measure connectivity. This involved the registration of consecutive images of Haematoxylin and Eosin (H&E) and Cytokeratin (CK) antibody stained tissue sections. Good examples of epithelial overlap were selected from these and were included within a pair of training images. The use of these with the Consensus Principal Component Analysis (CPCA) algorithm enabled for epithelial specific loadings to be identified.

These were used with the D Statistic to segment the epithelium from within three dimensional (3D) volumes of 13 cases of EM and 9 samples of cancer. The number of 26 3D connected components were recorded and used to test the null hypothesis that the cancer cases were as connected as the EM samples. In cancer, a higher number of connected components will be found because its submucosal glandular tissue has formed due to metastasis. In EM samples, the number of components will be lower as the epithelial islands remain in continuity from where they were misplaced.

However, analysis by the 99% confidence limits of the Mann Whitney U Statistic found no significant differences and so the null hypothesis could not be rejected. Although, no discrimination was brought about, many methodologies were investigated that had use within other studies. For example, image stitching algorithms [123] along with intermodal registration techniques [29] facilitated the development of digital staining functions and were used later with IR images.

It was thought that no differences were found because the adenocarcinomas were early stage cancers. Therefore, the epithelial islands are unlikely to have completely disassociated from the mucosal surface. However, other experimental design factors could have also impacted the results. Although, the movies contained within Electronic Appendix A show that the cancerous pathologies epithelium is as connected as the EM samples.

### 6.2 Immunohistochemistry (IHC)

Other diagnostic adjuncts suggested in the literature include staining with antibodies [9], [11–13]. It has been suggested that these can differentiate EM from cancer but further investigations found that only Collagen IV staining provided comparable results [13]. In this instance, Collagen IV basement membrane staining facilitated 100% differentiation between nine samples of EM and ten cases of cancer. The Mann Whitney U statistic found this to be statistically significant, $p < 0.001$, but the subjective assessment associated with this analysis is problematic.

This is because in this analysis not all of a samples glandular tissue was assessed and the opinion of others may differ to this investigations results. However, this antibody displays much potential and its further use with automated algorithms will mean that it may have application within the clinical arena.

Also in the future, the staining characteristics of a larger cohort of samples should also be investigated. The clustering of these samples within new HCA models could facilitate the identification of EM specific staining regimes.

### 6.3 Infrared (IR) Spectroscopy

The subjectivity of IHC analysis meant that other objective techniques were sought. Therefore, Infrared (IR) spectroscopy was used to probe the vibrational modes of mucinous compounds and epithelial tissue. Specific histological components were chosen because the retention of spectra relating to other features will affect multivariate analysis.

In order for such spectra to be derived, segmentation algorithms were required. In the first instance, mucin was identified using Principal Component Analysis (PCA) and the D Statistic. This allowed mucinous spectra to be collected. It was thought, that the further inclusion of these within a PCA fed Linear Discriminant Analysis (LDA) model would facilitate discrimination. However, it was noted that peaks related to paraffin dominated the analysis.

Therefore, ways of nullifying the contribution of this compound were investigated, such as by the Extended Multiplicative Scatter Correction (EMSC) algorithm [100], [104]. However, results suggested that the best way to treat this compound was to delete the regions of the IR spectrum that were related to it and this was implemented in all further analysis.

The inclusion of the processed spectra within a PCA fed LDA model provided no meaningful discrimination and so it was determined that mucin was not a good target

for IR. Another problem associated with mucin was that the pools did not always survive histological processing. Therefore, probing this compound is likely to have a limited clinical impact.

As an alternative, IR was used to measure individual glands that were contained within tissue sections. This involved a pathologist identifying pathology specific epithelial regions on images of standardly stained tissue sections. These were further non rigidly registered to low resolution IR maps, taken of consecutive sections. This precise alignment allowed IR images of the original identified regions to be acquired.

Although, segmentation techniques were still required so that epithelial specific spectra could be collected. Various methodologies were investigated but none were found that were fit for purpose and this was attributed to the spectrums poor Signal to Noise Ratios (SNR). The poor SNRs were surprising given that a high number of coadded scans were used but it was thought that this could be related to how the samples were processed.

Therefore, whole images of tissue sections were acquired using IR. In this instance, segmentation was achieved by using CPCA with CK and H&E images that had been registered to high resolution IR maps. However, some of the standardly stained tissue sections were destroyed during processing and this meant that others further down the block had to be used. This affected the reliable overlap of features and meant that this technique could not be used in the reliable derivation of epithelial specific spectra.

However, linear and non linear registration allowed some images of tissue samples to accurately align. The relation of these images PCA scores via linear regression enabled the development of H&E and E-Cadherin digital staining functions. These allowed for over 150,000 and 70,000 spectra to be derived from the EM and cancer cases respectively.

When these were included within a cross validated multivariate model both pathologies were correctly classified 80% of the time and this result is comparable to the current literature [84], [85], [94]. This is the first instance where IR maps have been visualised in this way and it is preferable to other methods as complete confidence can be asserted that epithelial specific spectra were being collected. In this instance, phospholipid peaks, present at 1740$cm^{-1}$ wavenumbers, were important in the characterisation of EM samples and this could be indicative of the fluidity of the cancerous cells. However, more work is required.

### 6.4 Future Work

The lack of differentiation brought about by 3D volume rendering could be due to missing tissue sections. When a samples tissue block is processed it is estimated that up to 30% of it could be lost [27]. Therefore, ways of estimating the appearance of these is required and interpolation algorithms could be used for this.

One example of these is the Kalman filter [76] and this allows the location of the glands from within the missing sections to be estimated. However, using the Kalman filter directly would require that the epithelial bodies contained within volumes be fairly independent of one another. For example, if an epithelial body changes path quickly within the missing sections of a 3D volume then the Kalman filter might find it difficult to predict its true path. In worst case scenarios, the filter might join two regions together that would have otherwise been disconnected. This is explained in more detail in Figure 6.1.

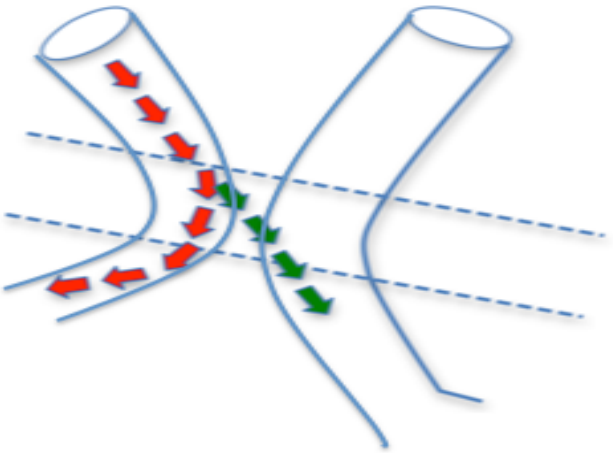


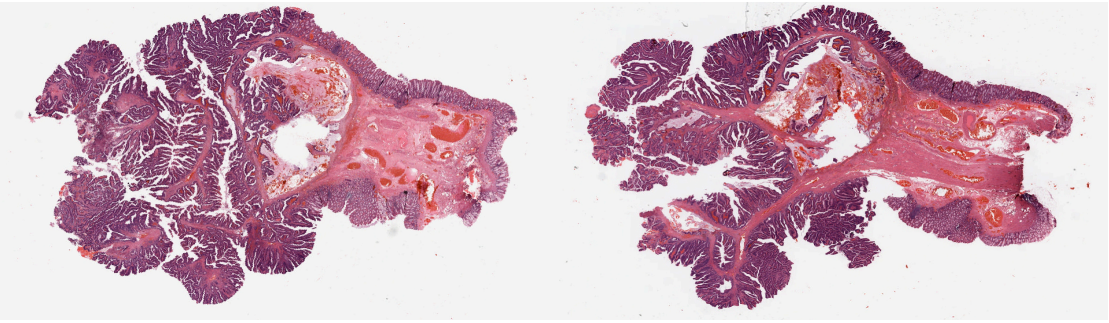
**Figure 6.1.** The possible false connectivity introduced by the Kalman filter. The dashed blue lines represent where there is missing tissue and the red arrows indicate this glandular tissues true path within a 3D volume. The green arrows represent the predicted path for the glandular tissue by the Kalman filter.

In the above figure, the dashed blue lines represent missing tissue sections and the green arrows indicate the predicted path for the epithelium by the linear Kalman filter. It can be seen that this will not follow the true path of the glandular tissue, indicated by the red arrows. Therefore, complex parameters will have to be used with this filter and this will make it both cumbersome and computationally inefficient.

Improvements to the experimental design used in the 3D volume rendering study might also help with analysis. Currently, only one out of the many tissue blocks available for EM samples was used and the epithelium within this might have already disassociated from the surface. Therefore, by using more of the tissue block in the

volume rendering process, a better representation of the continuity of epithelial structures will be brought about.

In this investigation, multiple tissue blocks were cut for one sample but the accurate registration of images sourced from the beginning and end of these blocks proved difficult. This is evident from Figure 6.2 A & B where it can be seen that the loss of tissue [27] has prevented accurate registrations.



A B

**Figure 6.2.** The loss of tissue that is associated with sectioning a tissue block. A) This is an image of the end tissue section from one block of a sample. B) This is the beginning tissue section cut from another block of the same sample. It is evident that there are stark differences in appearance between these images.

Interpolation could be used again to estimate how the epithelial features between the images of the different blocks are connected. This might help with determining how glandular structures are connected but will increase the computational complexity associated within this method.

Other ways of analysing 3D volumes should be considered [10], [58]. For example, investigating how the epithelial structures grow in 3D has been proven to be helpful elsewhere in the literature [10].

Technologies that circumvent some of the issues relating to tissue folding, missing tissue sections and segmentation should also be used. Examples of possible techniques include photonic solutions such as optical projection tomography [81] and life time imaging systems [181]. In these methods registration is not required and the spectroscopic information contained within the volumes can be used to segment features. Although, there is little evidence of their use with paraffin embedded tissue, some techniques require dyes and they are depth limited[81], [182].

For IHC analysis, automated algorithms should be implemented. For example, the Collagen IV staining of epithelial basement membranes could be quantified through segmentation algorithms. There are already well defined algorithms within the current

literature that enable the segmentation of membranous IHC stains and these could easily be implemented here [180].

Further to the development of automated methodologies a larger complement of antibodies should be tested. For example, survivin has been found to have a weak expression in both adenomatous and benign tissue but has a strong presence within adenocarcinomas[183].

Future IR work should consider the further development of the novel digital staining technique that this research developed. A high spectral and spatial resolution image of a whole tissue section should be acquired using a large number of coadded scans. This will be time consuming but will enable the accuracy of the digital staining to be improved. Only one image has to be processed in this way and this will allow for the staining characteristics of smaller areas measured from other samples to be predicted. However, the digital staining will be dependent on the tissue samples initial processing. In this study, it was noted that the mapping of single glands was problematic as the standard operating procedures used with some samples was poor.

Other options for digital staining include the development of a super stain that is highly specific for cancer. To achieve this, multiple contiguous IR maps and IHC stained slides, which have been stitched together, should be acquired. The registration of these will allow for their variances to be modelled and this will facilitate the production of digital staining functions. The cancer specific biochemistry that these are highlighting can be combined and this will allow a function to be produced that can accurately identify cancer.

This sort of analysis will be further aided by the incorporation of Raman spectroscopy with IR data. This can be achieved by registering images of each modality that have been acquired for a sample. Although, it should be noted that the use of Raman with paraffin embedded tissue is difficult and this technique has its own associated difficulties [92], [93], [184].

In conclusion, this study has highlighted the benefits of using intermodal registration [40] to bring about the discrimination of disease states. Novel methodologies have been presented that with further experimentation will enable for automated pathological adjuncts to be brought about. In addition to this, Collagen IV has been identified as a potential discriminatory biomarker and the future analysis of this stain with quantifiable means is justified.

# References


[1] D. David, W. Morson, and C. Basil, *Morson and Dawson's astrointestinal pathology*. Wiley-Blackwell, 2003, p. 567.

[2] B. Vogelstein, E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger, M. Leppert, Y. Nakamura, R. White, A. M. Smits, and J. L. Bos, "Genetic Alterations During Colorectal-Tumour Development.," *The New England Journal of Medicine*, vol. 319, no. 9, pp. 525–532, 1988.

[3] J. B. O'Connell, M. Maggard, and Y. Ko Clifford, "Colon Cancer Survival Rates With the New American Joint Committee on Cancer Sixth Edition Staging.," *Journal of National Cancer Institute*, vol. 96, no. 19, pp. 1420 – 1425, 2004.

[4] NHS BCSP, "Adenoma Surveillance.," 2009.

[5] "Guidelines from the Bowel Cancer Screening Programme Pathology Group: Reporting Lesions in the NHS Bowel Cancer Screening Programme.," 2007.

[6] J. Ferlay, P. Autier, M. Boniol, M. Heanue, M. Colombet, and P. Boyle, "Estimates of the Cancer Incidence and Mortality in Europe in 2006.," *Annals of oncology*, vol. 18, no. 3, pp. 581–92, Mar. 2007.

[7] "Cancer incidence for common cancers - UK statistics." [Online]. Available: http://info.cancerresearchuk.org/cancerstats/incidence/commoncancers/. [Accessed: 01-May-2010].

[8] F. L. Greene, "Epithelial Misplacement in Adenomatous Polyps of the Colon and Rectum," *Cancer*, vol. 33, no. 1, pp. 206–217, Jan. 1973.

[9] J. Mueller, E. Mueller, E. Arras, B. Bethke, and M. Stolte, "Stromelysin-3 Expression in Early (pT1) Carcinomas and Pseudoinvasive Lesions of the Colorectum.," *Virchows Archiv*, vol. 430, no. 3, p. 213:219, 1997.

[10] T. Tanizawa, T. Seki, M. Nakano, and R. Kamiyama, "Pseudoinvasion of the Colorectal Polypoid Tumors: Serial Section Study of Problematic Cases.," *Pathology international*, vol. 53, no. 9, pp. 584–90, Sep. 2003.

[11] T. P. Hansen, C. Fenger, and O. Kronborg, "The Expression of p53, Ki-67 and Urokinase Plasminogen Activator Receptor in Colorectal Adenomas with True Invasion and Pseudoinvasion.," *APMIS : acta pathologica, microbiologica, et immunologica Scandinavica*, vol. 107, no. 7, pp. 689–94, Jul. 1999.

[12] R. K. Yantiss, H. Goldman, and R. D. Odze, "Hyperplastic Polyp with Epithelial Misplacement (Inverted Hyperplastic Polyp): A Clinicopathologic and Immunohistochemical Study of 19 Cases.," *Modern pathology*, vol. 14, no. 9, pp. 869–75, Oct. 2001.

[13] R. K. Yantiss, M. Bosenberg, D. Antonioli, and R. Odze, “Utility of MMP-1, p53, E-Cadherin and Collagen IV Immunohistochemical Stains in the Differential Diagnosis of Adenomas with Misplaced Epithelium verus Adenomas with Invasive Adenocarcinoma.,” *The American journal of surgical pathology*, vol. 26, no. 2, pp. 206–215, 2002.

[14] C. Edo de Bock and Y. Wang, “Clinical Significance of Urokinase-type Plasminogen Activator Receptor (uPAR) Expression in Cancer.,” *Medicinal research reviews*, vol. 24, no. 1, pp. 13–39, Jan. 2004.

[15] P. Basset, J. Bellocq, C. Wolf, I. Stoll, P. Hutin, J. Limacher, O. Podhajcer, P. Chenard, M. Rio, and P. Chambon, “A Novel Metalloproteinase Gene Specifically Expressed in Stromal Cells of Breast Carcinomas.,” *Nature*, vol. 348, pp. 699–704, 1990.

[16] R. Visser, J. W. Arends, I. M. Leigh, and F. T. Bosman, “Patterns and Composition of Basement Membranes in Colon Adenomas and Adenocarcinomas.,” *Journal of pathology*, vol. 170, no. 3, pp. 285–290, 1993.

[17] R. Bertorelle, G. Esposito, C. Belluco, L. Bonaldi, A. Del Mistro, D. Nitti, and M. Lise, “p53 Gene Alterations and Protein Accumulation in Colorectal Cancer.,” *Journal of Clinical pathology: Molecular pathology*, vol. 49, no. 2, pp. 85–90, 1996.

[18] G. Giuseppe, O. Kandemir, D. Liu, M. Guida, S. Benvesito, T. Ruers, I. Benjamin, J. Northover, G. Stamp, I. Talbot, and M. Pignatelli, “Changes in E-Cadherin Immunoreactivity in the Adenoma-Carcinoma Sequence of the Large Bowel.,” *Virchows Archiv*, vol. 426, no. 2, pp. 149–154, 1995.

[19] C. M. Ellis, M. J. Dyson, T. J. Stephension, and E. L. Maltby, “HER2 Amplification Status in Breast Cancer: A Comparison between Immunohistochemical Staining and Fluorescence in situ Hybridisation using Manual and Automated Quantitative Image Analysis Scoring. techniques,” *Journal of clinical pathology*, vol. 58, no. 7, pp. 710–715, 2005.

[20] M. Cregger, A. J. Berger, and D. L. Rimm, “Immunohistochemistry and Quantitative Analysis of Protein Expression.,” *Archives of Pathology*, vol. 130, no. 7, pp. 1026–1030, 2006.

[21] M. S. Kim, T. Kim, S. Kong, S. Kwon, C. Y. Bae, J. Choi, C. Hwan, E. S. Lee, and J. Park, “Breast Cancer Diagnosis Using a Microfluidic Multiplexed Immunohistochemistry Platform.,” *Cancer*, vol. 5, no. 5, pp. 1–12, 2010.

[22] Z. Theodosiou, I. N. Kasampalidis, G. Livanos, M. Zervakis, I. Pitas, and K. Lyroudia, “Automated Analysis of FISH and Immunohistochemistry Images : A Review.,” *Cytometry*, vol. 71, no. 7, pp. 439–450, 2007.

[23] D. Turbin, S. Leung, M. C. U. Cheang, H. Kennecke, K. D. Montgomery, S. McKinney, D. O. Treaba, N. Boyd, L. C. Goldstein, S. Badve, A. M. Gown, M. van de Rijn, T. O. Nielsen, C. B. Gilks, and D. G. Huntsman, “Automated Quantitative Analysis of Estrogen Receptor Expression in Breast Carcinoma does not Differ from Expert Pathologist Scoring: A

Tissue Microarray Study of 3,484 Cases.," *Breast cancer research and treatment*, vol. 110, no. 3, pp. 417–26, Aug. 2008.

[24] S. Ourselin, "Reconstructing a 3D Structure from Serial Histological Sections.," *Image and Vision Computing*, vol. 19, no. 1–2, pp. 25–31, Jan. 2001.

[25] J. Streicher, W. Wolfgang, and M. Gerd, "External Marker Based Automatic Congruencing: A New Method of 3D Reconstruction from Serial Sections.," *The Anatomical Record*, vol. 248, pp. 583–602, 1997.

[26] K. Mosaliganti, "Registration and 3D Visualization of Large Microscopy Images.," in *Proceedings of SPIE*, 2006, vol. 6144, p. 61442V–61442V–12.

[27] S. Gaffling, J. Florian, V. Daum, M. Tauchi, and L. Elke, "Interpolation of Histological Slices by Means of Non-Rigid Registration.," in *Bildverarbeitng für die Medizin*, 2009, vol. 14, pp. 267–271.

[28] C. J. Solomon and T. P. Breckon, *Fundamentals of Digital Image Processing: A Practical Approach with Examples in Matlab.* Wiley-Blackwell, 2010.

[29] I. Arganda-carreras, C. O. S. Sorzano, and R. Marabini, "Consistent and Elastic Registration of Histological Sections using Vector-Spline Regularization.," in *CVAMIA'06 Proceedings of the Second ECCV international conference on Computer Vision Approaches to Medical Image Analysis*, 2006, vol. 4241, pp. 85–95.

[30] J. P. W. Pluim, J. B. A. Maintz, and M. Viergever, "Mutual-Information-Based Registration of Medical Images: A Survey.," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[31] S. Schaefer, T. McPhail, and J. Warren, "Image Deformation using Moving Least Squares.," *ACM Transactions on Graphics*, vol. 25, no. 3, p. 533, Jul. 2006.

[32] C. Tang, Y. Dong, and X. Su, "Automatic Registration Based on Improved SIFT for Medical Microscopic Sequence Images.," in *2008 Second International Symposium on Intelligent Information Technology Application*, 2008, pp. 580–583.

[33] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Mohee, and P. Quirke, "Colour Normalisation in Digital Histopathology Images.," in *Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy*, 2009, pp. 100–111.

[34] A. Pitiot, E. Bardinet, P. M. Thompson, and G. Malandain, "Piecewise Affine Registration of Biological Images for Volume Reconstruction.," *Medical image analysis*, vol. 10, no. 3, pp. 465–83, Jun. 2006.

[35] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints.," in *International Journal of Computer Vision*, 2004, vol. 60, no. 2, pp. 91–110.

[36] S. Saalfeld, A. Cardona, V. Hartenstein, and P. Tomancák, “As-rigid-as-possible Mosaicking and Serial Section Registration of Large ssTEM Datasets.,” *Bioinformatics*, vol. 26, no. 12, pp. 57–63, Jun. 2010.

[37] M. A. Fischler and R. C. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography.,” in *Communications of the ACM*, 1981, vol. 24, no. 6, pp. 381–395.

[38] N. M. Alpert, J. F. Bradshaw, D. Kennedy, and J. Correia, “The Principal Axes Transformation--a Method for Image Registration.,” *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 31, no. 10, pp. 1717–22, Oct. 1990.

[39] R. Sprengel, K. Rohr, and H. S. Stiehl, “Thin-Plate Spline Approximation for Image Registration.,” in *Proceedings of 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1996, vol. 3, pp. 1190–1191.

[40] W. R. Crum, “Non-rigid Image Registration: Theory and Practice.,” *British Journal of Radiology*, vol. 77, pp. 140–153, Dec. 2004.

[41] D. Magee, D. Treanor, and P. Quirke, “A New Image Registration Algorithm with Application to 3D Histopathology,” in *Proc. Microscopic Image Analysis with Applications in Biology*, 2008.

[42] N. Roberts, M. Derek, S. Yi, K. Brabazon, M. Shires, D. Crellin, N. M. Orsi, R. Quirke, P. Quirke, and D. Treanor, “Toward Routine Use of 3D Histopathology as a Research Tool.,” *The American Journal of Pathology*, vol. 180, no. 5, pp. 1835–1842, 2012.

[43] P. J. Van Diest and S. Al-janabi, “Digital Pathology : Current Status and Future Perspectives.,” *Histopathology*, vol. 61, no. 1, pp. 1–9, 2011.

[44] E. D. Castro and C. Morandi, “Registration of Translated and Rotated Images Using Finite Fourier Transforms.,” *IEEE Transaction on pattern analysis and machine intelligence*, vol. 9, no. 5, pp. 700–703, 1987.

[45] H. P. Menon and K. . Narayanankutty, “Applicability of Non-Rigid Medical Image Registration using Moving Least Squares.,” *International Journal of Computer Applications*, vol. 1, no. 6, pp. 85–92, Feb. 2010.

[46] J. T. Kwak, S. M. Hewitt, S. Sinha, and R. Bhargava, “Multimodal Microscopy for Automated Histologic Analysis of Prostate Cancer.,” *BMC cancer*, vol. 11, no. 62, Jan. 2011.

[47] D. C. Fernandez, R. Bhargava, S. M. Hewitt, and I. W. Levin, “Infrared Spectroscopic Imaging for Histopathologic Recognition.,” *Nature biotechnology*, vol. 23, no. 4, pp. 469–74, Apr. 2005.

[48] D. Naumann, P. Lasch, M. Diem, and W. Ha, “Artificial Neural Networks as Supervised Techniques for FT-IR Microspectroscopic Imaging.,” *Journal of Chemometrics*, vol. 20, no. 5, pp. 209–220, 2006.

[49] X. Wang, L. Li, C. Hu, J. Qiu, Z. Xu, and Y. Feng, “A Comparative Study of Three CT and MRI Registration Algorithms in Nasopharyngeal Carcinoma.,” *Journal of applied clinical medical physics*, vol. 10, no. 2, p. 2906, Jan. 2009.

[50] C. Kendall, M. Isabelle, F. Bazant-Hegemark, J. Hutchings, L. Orr, J. Babrah, R. Baker, and N. Stone, “Vibrational Spectroscopy: A Clinical Tool for Cancer Diagnostics.,” *The Analyst*, vol. 134, no. 6, pp. 1029–45, Jun. 2009.

[51] J. J. Wood, C. Kendall, G. R. LLoyd, N. Shepherd, T. Cook, and N. Stone, “Infrared spectroscopy to estimate the gross biochemistry associated with different colorectal pathologies,” in *Proceedings of SPIE: Clinical and Biomedical Spectroscopy and Imaging II*, 2011, vol. 8087, p. 80870P–80870P–9.

[52] L. E. Boucheron, Z. Bi, N. R. Harvey, B. S. Manjunath, and D. L. Rimm, “Utility of Multispectral Imaging for Nuclear Classification of Routine Clinical Histopathology Imagery.,” *BMC Cell Biology*, vol. 8, 2007.

[53] X. Mei and F. Porikli, “Fast Image Registration via Joint Gradient Maximization: Application to Multimodal Data.,” in *Proceedings of SPIE: Electro-Optical and Infrared Systems*, 2006, vol. 6395, p. 63950P–63950P–5.

[54] S. G. Kong, “Multiscale Fusion of Visible and Thermal IR Images for Illumination-Invariant Face Recognition.,” *International Journal of Computer Vision*, vol. 71, no. 2, pp. 215–233, 2007.

[55] A. Tangherlini, A. Merla, and G. L. Romani, “Field-Warp Registration for Biomedical High-Resolution Thermal Infrared Images.,” in *Proceedings of the 28th IEEE*, 2006, pp. 961–964.

[56] H. Abdelmunim, “Elastic Shape Registration using an Incremental Free Form Deformation Approach with the ICP Algorithm.,” in *Canadian Conference on Computer and Robot Vision (CRV)*, 2011, pp. 212–218.

[57] M. Tim, *Computer Vision and Image Processing.* Palgrave Macmillan Limited, 2003, pp. 47–98.

[58] J. Einenkel, J.-P. Kuska, L.-C. Horn, N. Wentzensen, M. Höckel, and U.-D. Braumann, “Combined Three-Dimensional Microscopic Visualisation of Tumour-Invasion Front of Cervical Carcinoma.,” *The lancet oncology*, vol. 7, no. 8, p. 698, Aug. 2006.

[59] C. E. H. Berger, J. A. de Koeijer, W. Glas, and H. T. Madhuizen, “Color Separation in Forensic Image Processing.,” *Journal of forensic sciences*, vol. 51, no. 1, pp. 100–2, Jan. 2006.

[60] A. Ruifrok and D. Johnston, “Quantification of Histochemical Staining by Color Deconvolution.,” *Anal Quant Cytol Histol.*, vol. 4, pp. 291–299, 2001.

[61] S. Egmentation, D. L. Pham, C. Xu, and J. L. Prince, “Current Methods in Medical Image Segmentation,” *Annual Reviews of Biomedical Engineering*, vol. 02, pp. 315–337, 2000.

[62] T. F. Chan and L. Vese, “Active Contours Without Edges.,” *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 10, no. 2, pp. 266–77, Jan. 2001.

[63] J. Malcolm, Y. Rathi, A. Yezzi, and A. Tannenbaum, “Fast Approximate Surface Evolution in Arbitrary Dimension.,” in *Proceedings of SPIE: Medical Imaging 2008: Image Processing*, 2008, vol. 6914, pp. 1–9.

[64] J. C. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms.,” in *Advanced Applications in Pattern Recognition*, Springer, 1981, pp. 1–26.

[65] A. Hafiane, F. Bunyak, and K. Palaniappan, “Fuzzy Clustering and Active Contours for Histopathology Image Segmentation and Nuclei Detection.,” *Lecture Notes in Comput Science*, vol. 5259, pp. 903–914, 2008.

[66] B. Chen, M. Mete, and S. Kockara, “Parameter-Free Multi-level Fuzzy C-Means Clustering for Unsupervised Structure Detection in Histological Images.,” in *SDPS 2010 Transformative Systems Conference*, 2010, pp. 1–5.

[67] S. Naik, S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi, “Gland Segmentation and Computerized Gleason Grading of Prostate Histology by Integrating Low- , High-level and Domain Specific Information.,” in *Proceedings of 2nd Workshop on Microsopic Image Analysis with Applications in Biology*, 2007, pp. 1–8.

[68] J. Zhang and T. Tan, “Brief Review of Invariant Texture Analysis Methods,” *Pattern Recognition*, vol. 35, no. 3, pp. 735–747, Mar. 2002.

[69] D. A. Clausi and M. E. Jernigan, “Designing Gabor Filters for Optimal Texture Separability.,” *Pattern Recognition*, vol. 33, no. 11, pp. 1835–1849, 2000.

[70] S. Doyle, M. Hwang, K. Shah, A. Madabhushi, M. Feldman, and J. Tomaszeweski, “Automated Grading of Prostate Cancer using Architectural and Textural Image Features.,” in *Biomedical Imaging: From Nano to Macro*, 2007, vol. 4th IEEE I, pp. 1284–1287.

[71] K. M. M. Rama and P. Shah, “Statistical Analysis of Textural Features for Improved Classification of Oral Histopathological Images.,” *Journal of Medical Systems*, vol. 36, pp. 865–881, 2012.

[72] B. Alberts, A. Johnson, and J. Lewis, *Molecular Biology of the Cell.* 2002.

[73] I. B. Buchwalow and B. Werner, *Immunohistochemistry: Basics and Methods.* 2010.

[74] K. Mosaliganti, L. Cooper, R. Sharp, R. Machiraju, G. Leone, K. Huang, and J. Saltz, “Reconstruction of Cellular Biological Structures from Optical Microscopy Data.,” *IEEE transactions on visualization and computer graphics*, vol. 14, no. 4, pp. 863–76, 2008.

[75] R. C. Tam and A. Fournier, “Image Interpolation using Unions of Spheres.,” *The Visual Computer*, vol. 14, pp. 401–414, 1998.

[76] T. Behrens, K. Rohr, and H. S. Stiehl, “Robust Segmentation of Tubular Structures in 3-D Medical Images by Parametric Object Detection and Tracking.,” *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*, vol. 33, no. 4, pp. 554–61, Jan. 2003.

[77] R. M. Cothren, R. Richards-Kortum, V. M. Sivak, M. Fitsmaurice, R. P. Rava, G. A. Boyce, M. Doxtader, T. B. Blackman, G. B. Hayes, M. S. Feld, and R. E. Petra, “Gastrointestinal Tissue Diagnosis by Laser-Induced Fluroescene Spectroscopy at Endoscopy.,” *Gastrointenstinal Endoscopy*, vol. 36, no. 2, pp. 105–111, 1990.

[78] L. P. Hariri, A. R. Tumlinson, D. G. Besselsen, U. Utzinger, E. W. Gerner, and J. K. Barton, “Endoscopic Optical Coherence Tomography and Laser-Induced Fluorescence Spectroscopy in a Murine Colon Cancer Model.,” *Lasers in surgery and medicine*, vol. 38, no. 4, pp. 305–13, Apr. 2006.

[79] L. M. Almond, J. Hutchings, N. Shepherd, H. Barr, N. Stone, and C. Kendall, “Raman Spectroscopy: a Potential Tool for Early Objective Diagnosis of Neoplasia in the Oesophagus.,” *Journal of biophotonics*, vol. 4, no. 10, pp. 685–95, Oct. 2011.

[80] P. R. Pfau, M. V Sivak, A. Chak, M. Kinnard, R. C. K. Wong, G. Isenberg, J. Izatt, A. Rollins, and V. Westphal, “Criteria for the Diagnosis of Dysplasia by Endoscopic Optical Coherence Tomography.,” *Gastrointestinal endoscopy*, vol. 58, no. 2, pp. 196–202, Aug. 2003.

[81] H. W. Detrich III, M. Westerfield, and L. I. Zon, “Optical Projection Tomography for Spatio-Temporal Analysis in the Zebrafish.,” *Methods in Cell Biology*, vol. 76, pp. 37–50, 2004.

[82] B. R. Wood, K. R. Bambery, C. J. Evans, M. Quinn, and D. McNaughton, “A Three-Dimensional Multivariate Image Processing Technique for the Analysis of FTIR Spectroscopic Images of Multiple Tissue Sections.,” *BMC medical imaging*, vol. 6, no. 12, Jan. 2006.

[83] S. E. Reichenbach, X. Tian, R. Lindquist, Q. Tao, A. Henderson, and J. C. Vickerman, “Interactive Spatio-Spectral Analysis of Three-Dimensional Mass-Spectral (3DxMS) Chemical Images.,” *Surface and Interface Analysis*, vol. 43, no. 1–2, pp. 529–534, Jan. 2011.

[84] S. Argov, J. Ramesh, A. Salman, I. Sinelnikov, J. Goldstein, H. Guterman, and S. Mordechai, “Diagnostic Potential of Fourier-Transform Infrared Microspectroscopy and Advanced Computational Methods in Colon Cancer Patients.,” *Journal of biomedical optics*, vol. 7, no. 2, pp. 248–54, Apr. 2002.

[85] J. Ramesh, A. Salman, S. Mordechai, S. Argov, J. Goldstein, and I. Sinelnikov, “FTIR Microscopic Studies on Normal , Polyp , and Malignant Human Colonic Tissues.,” *Subsurface Sensing Technologies and Applications*, vol. 2, no. 2, pp. 99–117, 2001.

[86] C. Krafft, G. Steiner, C. Beleites, and R. Salzer, “Disease Recognition by Infrared and Raman Spectroscopy.,” *Journal of biophotonics*, vol. 28, no. 1, pp. 13–28, 2009.

[87] P. D. Lewis, K. E. Lewis, R. Ghosal, S. Bayliss, A. J. Lloyd, J. Wills, R. Godfrey, P. Kloer, and L. J. Mur, “Evaluation of FTIR Spectroscopy as a Diagnostic Tool for Lung Cancer using Sputum.,” *BMC cancer*, vol. 10, no. 1, p. 640, Jan. 2010.

[88] A. Travo, O. Piot, R. Wolthuis, C. Gobinet, M. Michel, J. Bara, F. Marie-Elisabeth, and J. Pierre, “IR Spectral Imaging of Secreted Mucus: A New Promising Tool for the Histopathological Recognition of Human Colon Adenocarcinomas.,” *Histopathology*, vol. 56, no. 7, pp. 921–931, 2010.

[89] B. Bird, K. Bedrossian, L. Nora, M. Romeo, and M. Diem, “Detection of Breast Micro-Metastases in Auxillary Lymph Nodes by Infrared Micro-Spectral Imaging.,” *Analyst*, vol. 134, no. 6, pp. 1067–1076, 2009.

[90] P. Lasch, W. Haensch, D. Naumann, and M. Diem, “Imaging of Colorectal Adenocarcinoma using FT-IR Microspectroscopy and Cluster Analysis.,” *Biochimica et biophysica acta*, vol. 1688, no. 2, pp. 176–86, Mar. 2004.

[91] P. Siebert and F. Hildebrandt, *Vibrational Spectroscopy in Life Science.* Wiley, 2008, pp. 70–75.

[92] C. Krafft, D. Codrich, G. Pelizzo, and V. Sergo, “Raman and FTIR Microscopic Imaging of Colon Tissue: A Comparative Study.,” *Journal of biophotonics*, vol. 1, no. 2, pp. 154–69, May 2008.

[93] J. Hutchings, C. Kendall, B. Smith, N. Shepherd, H. Barr, and N. Stone, “The Potential for Histological Screening using a Combination of Rapid Raman Mapping and Principal Component Analysis.,” *Journal of biophotonics*, vol. 2, no. 1–2, pp. 91–103, Feb. 2009.

[94] A. Zwielly, S. Mordechai, and E. Bogomolny, “Advanced Statistical Techniques Applied to Comprehensive FTIR Spectra on Human Colonic Tissues.,” *Medical Physics*, vol. 37, no. 3, pp. 1047–1055, 2010.

[95] A. Salman, S. Argov, J. Ramesh, J. Goldstein, I. Sinelnikov, H. Guterman, and S. Mordechai, “FT-IR Microscopic Characterization of Normal and Malignant Human Colonic Tissues.,” *Cellular and Molecular Biology*, vol. 47, no. 22, pp. 159–166, 2001.

[96] J. Eaden, K. Abrams, H. McKay, H. Denley, and J. Mayberry, “Inter-Observer Variation Between General and Specialist Gastrointestinal Pathologists when Grading Dysplasia in Ulcerative Colitis.,” *The Journal of pathology*, vol. 194, no. 2, pp. 152–7, Jul. 2001.

[97] J. K. Turner, G. T. Williams, M. Morgan, M. Wright, and S. Dolwani, “Interobserver Agreement in the Reporting of Colorectal Polyp Pathology among Bowel Cancer Screening Pathologists in Wales.,” *Histopathology*, vol. 62, no. 6, pp. 916–24, May 2013.

[98] R. Bhargava and I. Levin, *Spectrochemical Analysis using Infrared Multichannel Detectors.* Blackwell, 2005, pp. 1–5.

[99] R. Melissa J et al, *Introduction to Spectral Imaging, and Applications to Diagnosis of Lymph Nodes. Vibrational Spectroscopy for Medical Diagnosis.* Wiley, 2008, pp. 6–12.

[100] P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke, and P. Gardner, "Resonant Mie Scattering (RMieS) Correction of Infrared Spectra from Highly Scattering Biological Samples.," *Analyst*, vol. 135, no. 2, pp. 268–277, 2010.

[101] B. Bird, M. Miljković, and M. Diem, "Two Step Resonant Mie Scattering Correction of Infrared Micro-Spectral Data: Human Lymph Node Tissue.," *Journal of biophotonics*, vol. 3, no. 8–9, pp. 597–608, Aug. 2010.

[102] E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait, "Combination of FTIR Spectral Imaging and Chemometrics for Tumour Detection from Paraffin-Embedded Biopsies.," *The Analyst*, vol. 133, no. 2, pp. 197–205, Mar. 2008.

[103] J. Pijanka, G. D. Sockalingum, A. Kohler, Y. Yang, F. Draux, G. Parkes, K.-P. Lam, D. Collins, P. Dumas, C. Sandt, D. G. van Pittius, G. Douce, M. Manfait, V. Untereiner, and J. Sulé-Suso, "Synchrotron-based FTIR Spectra of Stained Single Cells. Towards a Clinical Application in Pathology.," *Laboratory investigation; a journal of technical methods and pathology*, vol. 90, no. 5, pp. 797–807, May 2010.

[104] D. Sebiskveradze, E. Ly, C. Gobinet, O. Piot, M. Manfait, P. Jeannesson, and V. Vrabie, "From Preprocessing to Fuzzy Classification of IR Images of Paraffin Embedded Cancerous Skin Samples.," in *2009 First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2009, pp. 1–4.

[105] E. Ó Faoláin, M. B. Hunter, J. M. Byrne, P. Kelehan, M. McNamara, H. J. Byrne, and F. M. Lyng, "A Study Examining the Effects of Tissue Processing on Human Tissue Sections using Vibrational Spectroscopy.," *Vibrational Spectroscopy*, vol. 38, no. 1–2, pp. 121–127, Jul. 2005.

[106] Y. Oxaki and K. Murayama, "IR and Raman Spectroscopy and Chemometrics.," in *Infrared and raman spectroscopy of biological materials*, 2001, pp. 515–565.

[107] H. Martens and E. Stark, "Extended Multiplicative Signal Correction and Spectral Interference Subtraction: New Preprocessing Methods for Near Infrared Spectroscopy.," *Journal of pharmaceutical and biomedical analysis*, vol. 9, no. 8, pp. 625–35, Jan. 1991.

[108] G. Lloyd, R. Bereton, and J. Duncan, "Self Organising Maps for Distinugishing Polymer Groups using Thermal Response Curves Obtained by Dynamic Mechanical Analysis.," *Analyst*, vol. 133, no. 8, pp. 1046–1059, 2008.

[109] Y. Ajioka, H. Watanabe, and J. R. Jass, "MUC1 and MUC2 Mucins in Flat and Polypoid Colorectal Adenomas.," *Journal of clinical pathology*, vol. 50, no. 5, pp. 417–21, May 1997.

[110] R. Walsh and J. Jass, "Altered Mucin Expression in the Gastrointestinal Tract; A Review.," *Journal of Cellular and Molecular Medicine*, vol. 5, no. 3, pp. 327–351, 2001.

[111] Y. Ajioka, L. J. Allison, and J. R. Jass, “Significance of MUC1 and MUC2 Mucin Expression in Colorectal Cancer.,” *Journal of clinical pathology*, vol. 49, no. 7, pp. 560–4, Jul. 1996.

[112] Y. S Kim, “Altered Glycosylation of Mucin Glycoproteins in Colonic Neoplasia.,” *Journal of Cellular Biochemistry - Supplement*, vol. 16, pp. 91–96, 1992.

[113] A.-E. Biemer-Hüttmann, M. Walsh, M. McGuckin, A. Y, W. H, L. B. A, and J. J. R, “Immunohistochemical Staining Patterns of MUC1, MUC2, MUC4, and MUC5AC Mucins in Hyperplastic Polyps, Serrated Adenomas, and Traditional Adenomas of the Colorectum.,” *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society*, vol. 47, no. 8, pp. 1039–48, Aug. 1999.

[114] M. D. Walsh, M. A. Mcguckin, L. A. Simms, J. Young, B. A. Leggett, and J. R. Jass, “Mucin Core Protein Expression in Colorectal Cancers with High Levels of Microsatellite Instability Indicates a Novel Pathway of Morphogenesis.,” *Clinical Cancer Research*, vol. 6, pp. 1909–1916, 2000.

[115] X. D. Bu, “Altered Expression of MUC2 and MUC5AC in Progression of Colorectal Carcinoma.,” *World Journal of Gastroenterology*, vol. 16, no. 32, pp. 4089–4094, 2010.

[116] K. Rajpoot and N. Rajpoot, “SVM Optimization for Hyperspectral Colon Tissue Cell Classification.,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004 Lecture Notes in Computer Science*, 2004, pp. 829–837.

[117] K. Masood and N. Rajpoot, “Spatial Analysis for Colon Biopsy Classification from Hyperspectral Imagery.,” in *Annals of the BMVA*, 2008, vol. 2008, no. 4, pp. 1–16.

[118] K. Masood, N. Rajpoot, K. Rajpoot, and H. Qureshi, “Hyperspectral Colon Tissue Classification using Morphological Analysis.,” in *IEEE 2nd International Conference on Emerging technologies*, 2006, pp. 13–14.

[119] V. J. Dercksen, B. Cornelia, D. Stalling, G. Sabine, U. Seiffert, and H.-C. Hege, “Towards Automatic Generation of 3D Models of Biological Objects Based on Serial Sections.,” *Mathematics and Visualization*, vol. 1, pp. 3–25, 2008.

[120] D. Rueckert, C. Sonoda, D. Hayes, G. Hill, M. Leach, and D. J. Hawkes, “Nonrigid Registration using Free-Form Deformations: Application to Breast MR Images.,” *IEEE transactions on medical imaging*, vol. 18, no. 8, pp. 712–21, Aug. 1999.

[121] J. D. Cass, S. Varma, A. G. Day, W. Sangrar, A. B. Rajput, L. H. Raptis, J. Squire, Y. Madarnas, S. K. SenGupta, and B. E. Elliott, “Automated Quantitative Analysis of p53, Cyclin D1, Ki67 and pERK Expression in Breast Carcinoma Does Not Differ from Expert Pathologist Scoring and Correlates with Clinico-Pathological Characteristics.,” *Cancers*, vol. 4, no. 4, pp. 725–742, Jul. 2012.

[122] D. Carey, G. R. Lloyd, N. Shepherd, and N. Stone, “Correlating Histology and Spectroscopy to Differentiate Pathologies of the Colon.,” in *Proceedings of the 16th Conference on Medical Image and Understanding and Analysis.*, 2012, pp. 243–248.

[123] V. Rankov, “An Algorithm for Image Stitching and Blending.,” in *Proceedings of SPIE*, 2005, vol. 5701, pp. 190–199.

[124] S. Preibisch, S. Saalfeld, and P. Tomancak, “Globally Optimal Stitching of Tiled 3D Microscopic Image Acquisitions.,” *Bioinformatics Applications Note*, vol. 25, no. 11, pp. 1463–1465, 2009.

[125] J. Chappelow, J. E. Tomaszewski, M. Feldman, N. Shih, and A. Madabhushi, “Computerized Medical Imaging and Graphics HistoStitcher © : An interactive Program for Accurate and Rapid Reconstruction of Digitized Whole Histological Sections from Tissue Fragments.,” *Computerized Medical Imaging and Graphics*, vol. 35, no. 7–8, pp. 557–567, 2011.

[126] A. Kovnatsky, “Feature points in image, Keypoint extraction,” *Matlab Fileexchange*, 2010. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/29004-feature-points-in-image-keypoint-extraction.

[127] K. Mikolajczyk, “Scale & Affine Invariant Interest Point Detectors.,” *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, Oct. 2004.

[128] C. Harris, “A Combined Corner and Edge Detector.,” in *In Proc. of Fourth Alvey Vision Conference*, 1988.

[129] C. Kenney, M. Bober, and B. Manjunath, “RANSAC Toolbox for Matlab.” 2004.

[130] C. D. Kuglin and D. C. Hines, “The Phase Correlation Image Alignment Method.,” in *IEEE international Conference on Cybernetics and Society*, 1975, pp. 163–165.

[131] L. Pantanowitz, P. N. Valenstein, A. J. Evans, K. J. Kaplan, J. D. Pfeifer, D. C. Wilbur, L. C. Collins, and T. J. Colgen, “Review of the Current State of Whole Slide Imaging in Pathology.,” *Journal of Pathology Informatics*, vol. 2, no. 36, 2011.

[132] M. G. Rojo, G. B. García, and M. C. Vicente, “Techniques in Pathology Critical Comparison of 31 Commercially Available Digital Slide Systems in Pathology.,” *International Jpournal of Surgical Pathology.*, vol. 14, no. 4, pp. 285–305, 2003.

[133] J. Modersitzki, O. Schmitt, and S. Writz, “Registration of Histological Serial Sectionings.,” in *Mathematics in Industry*, 10th ed., Springer Berlin Heidelberg, 2006, pp. 63–80.

[134] J. C. Lagaris, J. A. Reeds, M. H. Wright, and P. E. Wright, “Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions.,” *Society for Industrial and Applied Mathematics*, vol. 9, no. 1, pp. 112–147, 1998.

[135] G. Lombardi, “Moving Least Squares.” 2006.

[136] J. Canny, “A Computational Approach to Edge Detection.,” *IEEE Transaction on pattern analysis and machine intelligence*, vol. 8, no. 6, pp. 679–698, 1986.

[137] J. A. Westerhuis, T. Kourti, and J. F. Macgregor, “Analysis of multiblock and hierarchical PCA and PLS models.,” *Journal of Chemometrics*, vol. 321, no. May, pp. 301–321, 1998.

[138] R. Brereton, *Chemometrics for Pattern Recognition.* Wiley-Blackwell, 2009, p. 243.

[139] Z. Movasaghi, S. Rehman, and I. ur Rehman, “Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues.,” *Applied Spectroscopy Reviews*, vol. 43, no. 2, pp. 134–179, Feb. 2008.

[140] K. Gough, M. Rak, A. Bookatz, M. Del Bigio, S. Mai, and D. Westaway, “Choices for Tissue Visualization with IR Microspectroscopy,” *Vibrational Spectroscopy*, vol. 38, no. 1–2, pp. 133–141, Jul. 2005.

[141] A. Savitzky and M. J. E. Golay, “Smooting and Differentiation of Data by Simplified Least Squares Procedures.,” *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[142] N. Stone, C. Kendall, N. Shepherd, P. Crow, and H. Barr, “Near-Infrared Raman Spectroscopy for the Classification of Epithelial Pre-Cancers and Cancers.,” *Journal of Raman Spectroscopy*, vol. 33, no. 7, pp. 564–573, 2002.

[143] M. Khanmohammadi, “Application of Linear Discriminant Analysis and Attenuated Total Reflectance Fourier Transform Infrared Microspectroscopy for Diagnosis of Colon Cancer.,” *Pathology & Oncology Research*, vol. 17, no. 2, pp. 435–441, 2011.

[144] N. Shepherd, H. J. Bussey, and J. R. Jass, “Epithelial Misplacement in Peutz-Jeghers polyps. A diagnostic pitfall.,” *The American journal of surgical pathology*, vol. 11, no. 10. pp. 743–9, Oct-1987.

[145] T. J. Byun, D. S. Han, S. B. Ahn, H. S. Cho, C. S. Eun, Y. C. Jeon, J. H. Sohn, and Y.-H. Oh, “Pseudoinvasion in an Adenomatous Polyp of the Colon Mimicking Invasive Colon Cancer.,” *Gut and liver*, vol. 3, no. 2, pp. 130–3, Jun. 2009.

[146] V. C. Petersen, L. Sheehan, R. L. Bryan, C. P. Armstrong, and N. Shepherd, “Misplacement of Dysplastic Epithelium in Peutz-Jeghers Polyps: The Ultimate Diagnostic Pitfall?,” *The American journal of surgical pathology*, vol. 24, no. 1, pp. 34–9, Jan. 2000.

[147] J. Dauguet, T. Delzescaux, F. Condé, J. F. Mangin, N. Ayache, P. Hantraye, and V. Frouin, “Three-Dimensional Reconstruction of Stained Histological Slices and 3D Non-Linear Registration with in-vivo MRI for Whole Baboon Brain.,” *Journal of neuroscience methods*, vol. 164, no. 1, pp. 191–204, Aug. 2007.

[148] J. L. Ostuni, J. A. Frank, R. L. Levin, and C. Decarli, “Correspondence of Closest Gradient Voxels - A Robust Registration Algorithm.,” *Journal of Magnetic Resonance Imaging*, vol. 7, no. 2, pp. 410–415, 1997.

[149] O. Schmitt, J. Modersitzki, S. Heldmann, S. Wirtz, and B. Fischer, “Image Registration of Sectioned Brains.,” *International Journal of Computer Vision*, vol. 73, no. 1, pp. 5–39, Sep. 2006.

[150] V. E. C. Ghaleh and A. Behrad, “Lip Contour Extraction using RGB Color Space and Fuzzy C-Means Clustering.,” in *2010 IEEE 9th International Conference on Cyberntic Intelligent Systems*, 2010, pp. 1–4.

[151] P. Makowski, T. S. Sørensen, S. V. Therkildsen, A. Materka, H. Stødkilde-Jørgensen, and E. M. Pedersen, “Two-Phase Active Contour Method for Semiautomatic Segmentation of the Heart and Blood Vessels from MRI Images for 3D Visualization.,” *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, vol. 26, no. 1, pp. 9–17, 2002.

[152] J. Shiozawa, M. Ito, T. Nakayama, M. Nakashima, S. Kohno, and I. Sekine, “Expression of Matrix Metalloproteinase-1 in Human Colorectal Carcinoma.,” *Modern pathology : an official journal of the United States and Canadian Academy of Pathology*, vol. 13, no. 9, pp. 925–33, Sep. 2000.

[153] O. A. Stefansson, J. G. Jonasson, K. Olafsdottir, H. Hilmarsdottir, G. Olafsdottir, M. Esteller, O. T. Johannsson, and J. E. Eyfjord, “CpG Island Hypermethylation of BRCA1 and Loss of pRb as Co-Occurring Events in Basal/Triple-Negative Breast Cancer.,” *Epigenetics : official journal of the DNA Methylation Society*, vol. 6, no. 5, pp. 638–49, May 2011.

[154] C. Conti, P. Ferraris, E. Giorgini, C. Rubini, S. Sabbatini, G. Tosi, J. Anastassopoulou, P. Arapantoni, E. Boukaki, and S. Konstadoudakis, “FT-IR Microimaging Spectroscopy: A Comparison between Healthy and Neoplastic Human Colon Tissues.,” *Journal of Molecular Structure*, vol. 881, no. 1–3, pp. 46–51, Jun. 2008.

[155] T. K. Jin, S. Saurabh, and B. Rohit, “A New Segmentation Framework for Infrared Spectroscopic Imaging using Frequent Pattern Mining.,” in *Biomedical Imaging: From Nano to Macro*, 2011, pp. 452–455.

[156] B. Bird, M. M. Romeo, J. Smith, N. Stone, M. George, and M. Diem, “Infrared Micro-Spectral Imaging: Distinction of Tissue Types in Auxillary Lymph Node Histology.,” *BMC Clinical Pathology*, vol. 8, no. 8, 2008.

[157] I. Brockhausen, “Mucin-type O-glycans in Human Colon and Breast Cancer: Glycodynamics and Functions.,” *EMBO reports*, vol. 7, no. 6, pp. 599–604, Jun. 2006.

[158] R. Bansil and S. T. Bradley, “Mucin Structure, Aggregation, Physiological Functions and Biomedical Applications.,” *Current Opinion in Colloid & Interface Science*, vol. 11, no. 2–3, pp. 164–170, Jun. 2006.

[159] M. Molaei, B. K. Mansoori, R. Mashayekhi, M. Vahedi, M. A. Pourhoseingholi, S. R. Fatemi, and M. R. Zali, “Mucins in Neoplastic Spectrum of Colorectal Polyps: Can they Provide Predictions?,” *BMC cancer*, vol. 10, no. 537, Jan. 2010.

[160] L. J. M. Mekenkamp, K. J. Heesterbeek, M. Koopman, J. Tol, S. Teerenstra, S. Venderbosch, C. J. Punt, and I. D. Nagtegaal, “Mucinous Adenocarcinomas: Poor Prognosis in Metastatic Colorectal Cancer.,” *European journal of cancer*, vol. 48, no. 4, pp. 501–509, Jan. 2012.

[161] K. Matsuda, T. Masaki, T. Watanabe, J. Kitayama, H. Nagawa, T. Muto, and Y. Ajioka, "Clinical Significance of MUC1 and MUC2 Mucin and p53 Protein Expression in Colorectal Carcinoma.," *Japanese journal of clinical oncology*, vol. 30, no. 2, pp. 89–94, Mar. 2000.

[162] D. W. Kufe, "Mucins in Cancer: Function, Prognosis and Therapy.," *Nature reviews. Cancer*, vol. 9, no. 12, pp. 874–85, Dec. 2009.

[163] R. S. Bresalier, Y. Niv, J. C. Byrd, Q. Y. Duh, N. W. Toribara, R. W. Rockwell, R. Dahiya, and Y. S. Kim, "Mucin Production by Human Colonic Carcinoma Cells Correlates with their Metastatic Potential in Animal Models of Colon Cancer Metastasis.," *The Journal of clinical investigation*, vol. 87, no. 3, pp. 1037–45, Mar. 1991.

[164] L. R. Sternberg, J. C. Byrd, C. K. Yunker, S. Dudas, V. K. Hoon, and R. S. Bresalier, "Liver Colonization by Human Colon Cancer Cells is Reduced by Antisense Inhibition of MUC2 Mucin Synthesis.," *Gastroenterology*, vol. 116, no. 2, pp. 363–71, Feb. 1999.

[165] P. Lasch, W. Wäsche, W. J. Mccarthy, G. Müller, and D. Naumann, "Imaging of Human Colon Carcinoma Thin Sections by FT-IR Microspectrometry.," in *Proc. SPIE 3257, Infrared Spectroscopy: New Tool in Medicine*, 1998, pp. 187–198.

[166] G. Socrates, *Infrared and Raman Characteristic Group Frequencies.* Wiley, 2001, pp. 339–355.

[167] A. Roque, I. Ponte, and P. Suau, "Secondary Structure of Protamine in Sperm Nuclei: An Infrared Spectroscopy Study.," *BMC structural biology*, vol. 11, no. 14, Jan. 2011.

[168] F. Lyng, E. Gazim, and P. Gardener, "Preparation of Tissues and Cells for Infrared and Raman Spectroscopy and Imaging.," in *Biomedical Applications of Synchrotron Infrared Microspectroscopy*, 2011, pp. 145–188.

[169] J. G. Kelly, M. N. Singh, H. F. Stringfellow, M. J. Walsh, J. M. Nicholson, F. Bahrami, K. M. Ashton, M. Pitt, P. L. Martin-Hirsch, and F. L. Martin, "Derivation of a Subtype-Specific Biochemical Signature of Endometrial Carcinoma using Synchrotron-Based Fourier-Transform Infrared Microspectroscopy.," *Cancer letters*, vol. 274, no. 2, pp. 208–17, Feb. 2009.

[170] S. Rehman, "FTIR and Raman Characteristic Peak Frequencies in Biological Studies.," in *Vibrational Spectroscopy for Tissue Analysis*, Taylor & Francis, 2012, pp. 266–294.

[171] J. Wang, Y.-J. Geng, B. Guo, T. Klima, B. N. Lal, J. T. Willerson, and W. Casscells, "Near-Infrared Spectroscopic Characterization of Human Advanced Atherosclerotic Plaques.," *Journal of the American College of Cardiology*, vol. 39, no. 8, pp. 1305–13, Apr. 2002.

[172] A. Hynes, D. Scott, A. Man, D. L. Singer, M. G. Sowa, and K.-Z. Liu, "Molecular Mapping of Periodontal Tissues using Infrared Microspectroscopy.," *BMC medical imaging*, vol. 5, no. 2, May 2005.

[173] C. Petibois, G. Gouspillou, K. Wehbe, J.-P. Delage, and G. Déléris, “Analysis of Type I and IV Collagens by FT-IR Spectroscopy and Imaging for a Molecular Investigation of Skeletal Muscle Connective Tissue.,” *Analytical and bioanalytical chemistry*, vol. 386, no. 7–8, pp. 1961–6, Dec. 2006.

[174] L. B. Mostaço-guidolin, L. Sayuri, and M. Ribeiro, “Molecular and Chemical Characterization by Fourier Transform Infrared Spectroscopy of Human Breast Cancer Cells with Estrogen Receptor Expressed and Not Expressed.,” *Journal of Spectrocopy*, vol. 24, no. 5, pp. 501–510, 2010.

[175] Q.-B. Li, Z. Xu, N.-W. Zhang, L. Zhang, F. Wang, L.-M. Yang, J.-S. Wang, S. Zhou, Y.-F. Zhang, X.-S. Zhou, J.-S. Shi, and J.-G. Wu, “In vivo and in situ Detection of Colorectal Cancer using Fourier Transform Infrared Spectroscopy.,” *World journal of gastroenterology : WJG*, vol. 11, no. 3, pp. 327–30, Jan. 2005.

[176] B. Rigas and P. T. T. Wong, “Human Colon Adenocarcinoma Cell Lines Display Infrared Spectroscopic Features of Malignant Colon Tissues Human Colon Adenocarcinoma Cell Lines Display Infrared Spectroscopic Features of Malignant Colon Tissues.,” *Cancer Research*, vol. 52, no. 1, pp. 84–88, 1992.

[177] S. Duraipandian, “Real-Time Raman Spectroscopy for in vivo , Online Gastric Cancer Diagnosis During Clinical Endoscopic Examination.,” *Journal of Biomedical Optics*, vol. 17, no. 8, 2012.

[178] G. I. Dovbeshko, N. Y. Gridina, E. B. Kruglova, and O. P. Pashchuk, “FTIR Spectroscopy Studies of Nucleic Acid Damage.,” *Talanta*, vol. 53, no. 1, pp. 233–46, Oct. 2000.

[179] N. Stone, “Near-Infrared Raman Spectroscopy for Detection and Classification of Gastrointestinal Disease.,” in *Proceedings of SPIE*, 2002, vol. 4614, pp. 117–126.

[180] E. Ficarra, S. D. Cataldo, A. Acquaviva, and E. Macii, “Automated Segmentation of Cells With IHC Membrane Staining.,” *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1421–1429, 2011.

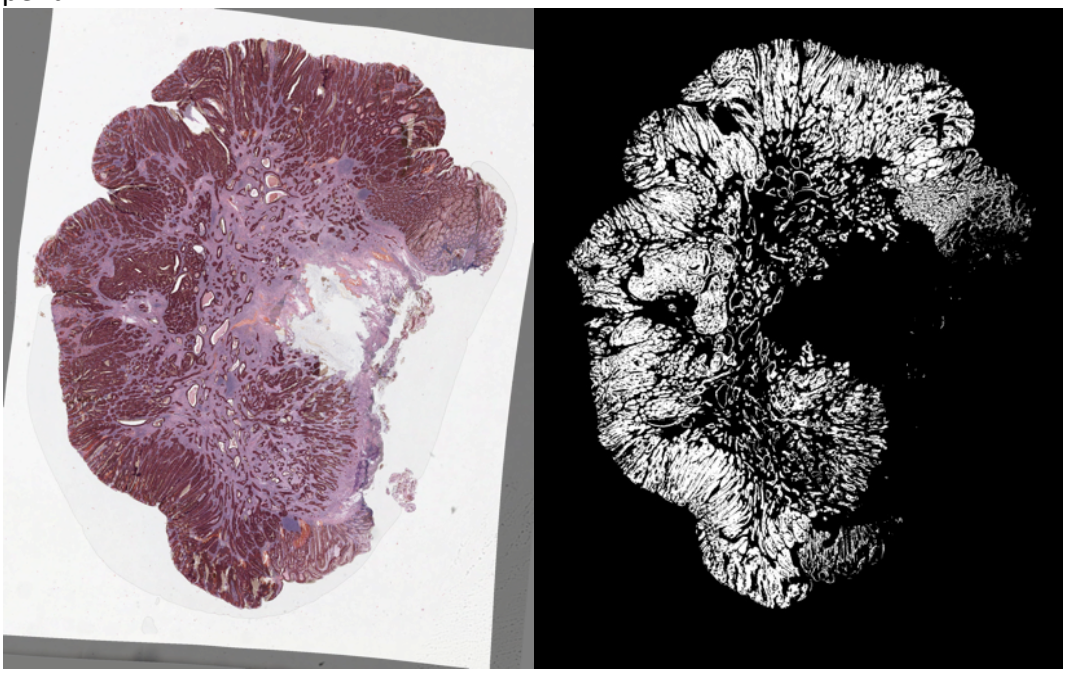
[181] S. M. Ameer-beg, P. R. Barber, R. J. Hodgkiss, R. J. Locke, R. G. Newman, and G. M. Tozer, “Application Of Multiphoton Steady State and Lifetime Imaging to Mapping of Tumour Vascular Architecture in vivo.,” in *Proc. SPIE 4620, Multiphoton Microscopy in the Biomedical Sciences II*, 2002, pp. 85–95.

[182] K. M. Short, M. J. Hodson, and I. M. Smyth, “Tomographic Quantification of Branching Morphogenesis and Renal Development.,” *Kidney international*, vol. 77, no. 12, pp. 1132–9, Jun. 2010.

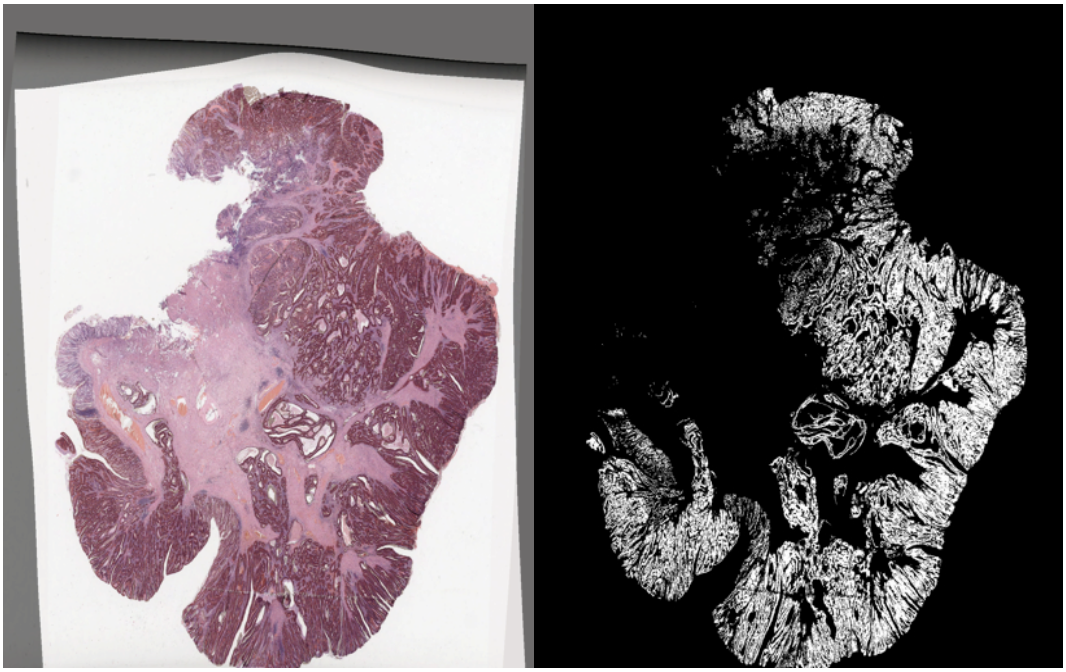
[183] J. M. Hernandez, J. M. Farma, D. Coppola, A. Hakam, W. J. Fulp, D.-T. Chen, E. M. Siegel, T. J. Yeatman, and D. Shibata, “Expression of the Antiapoptotic Protein Survivin in Colon Cancer.,” *Clinical colorectal cancer*, vol. 10, no. 3, pp. 188–93, Sep. 2011.

[184] S. Candefjord and O. A. Lindahl, “Effects of Snap-freezing and Near-Infrared Laser Illumination on Porcine Prostate Tissue as Measured by Raman Spectroscopy.,” *Analyst*, vol. 134, no. 6, pp. 1815–1821, 2009.

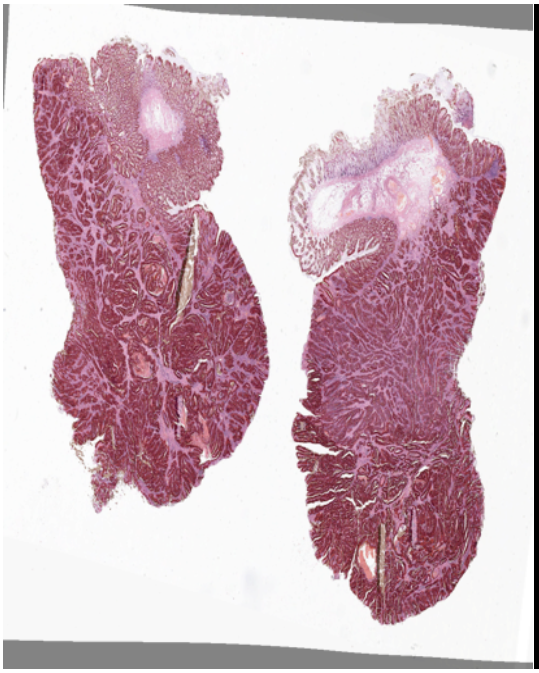
Appendix A:



A B



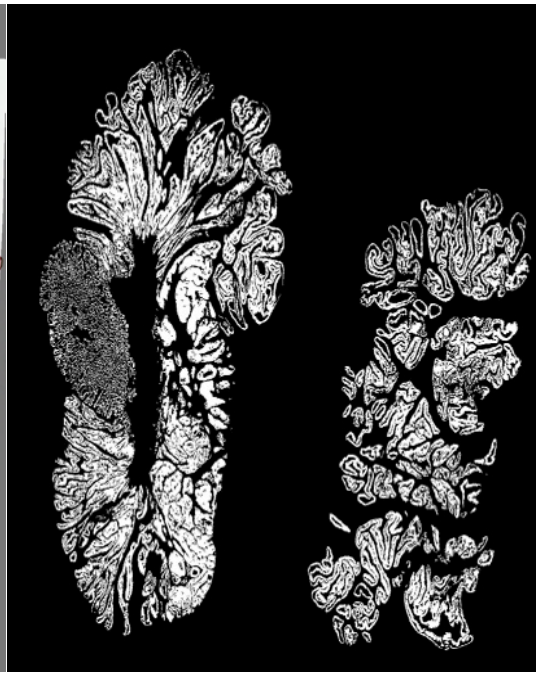
C D



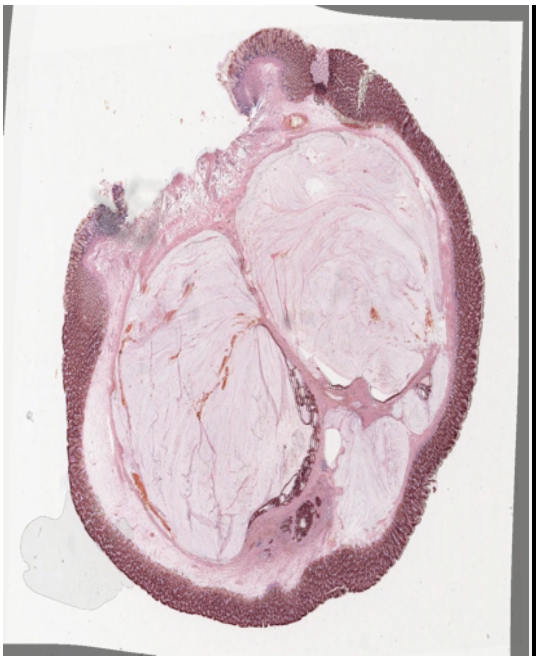
E

F

G

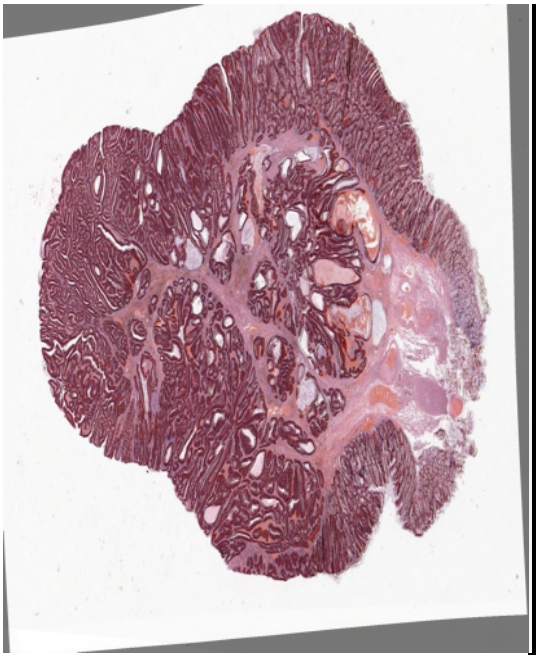


H

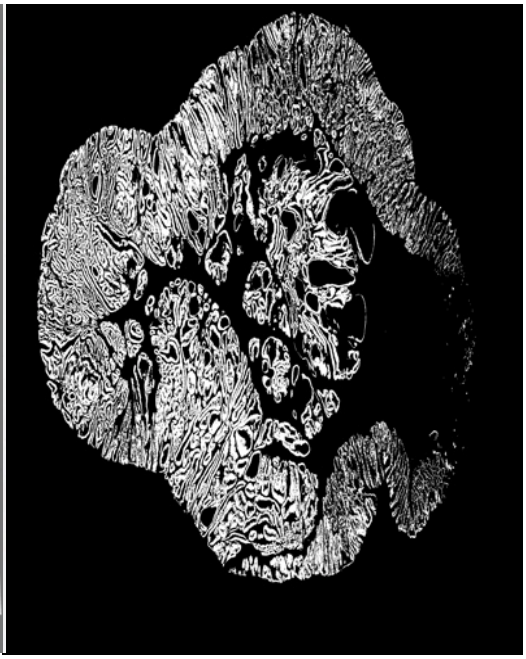


I

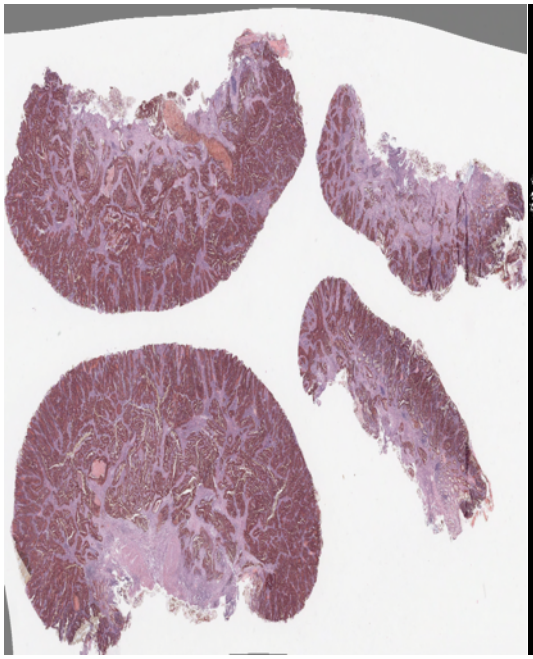
J



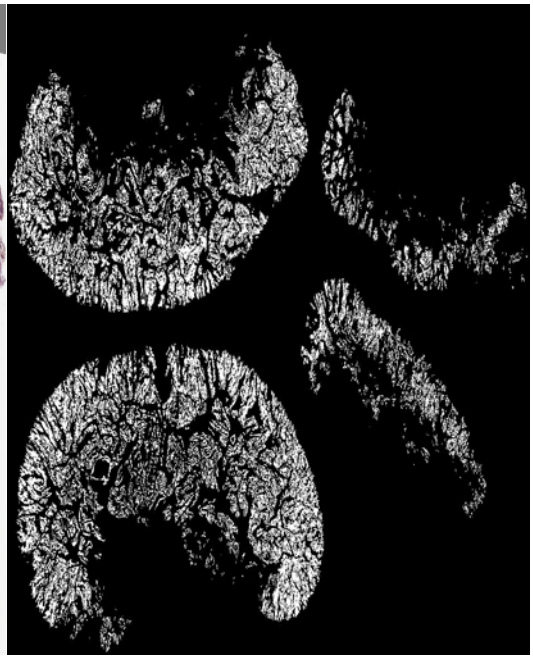
K



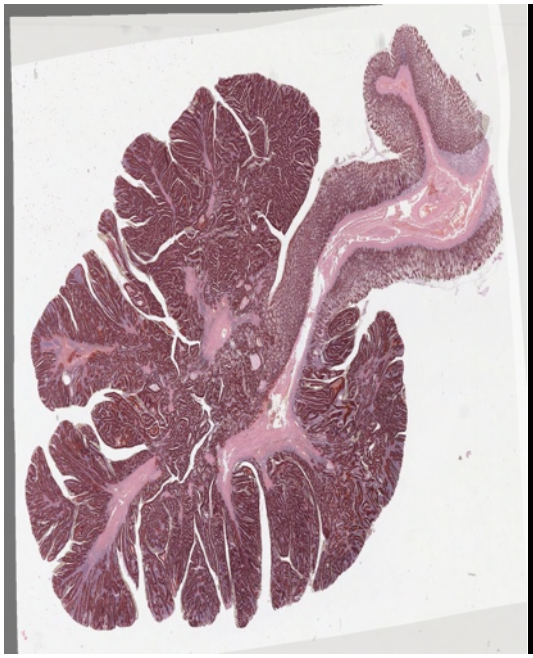
L



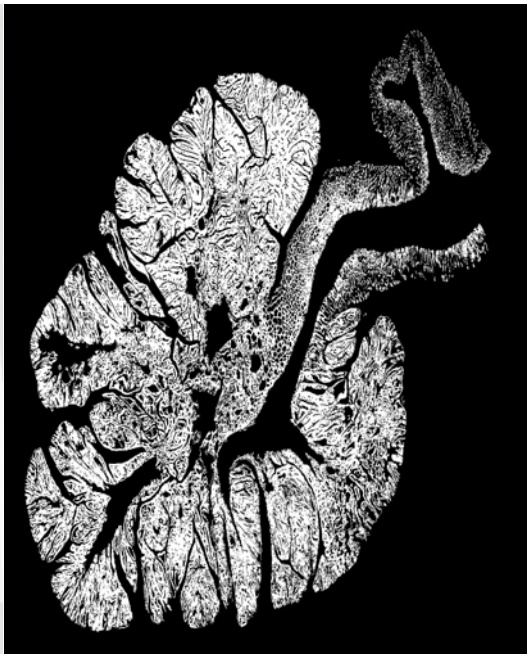
M



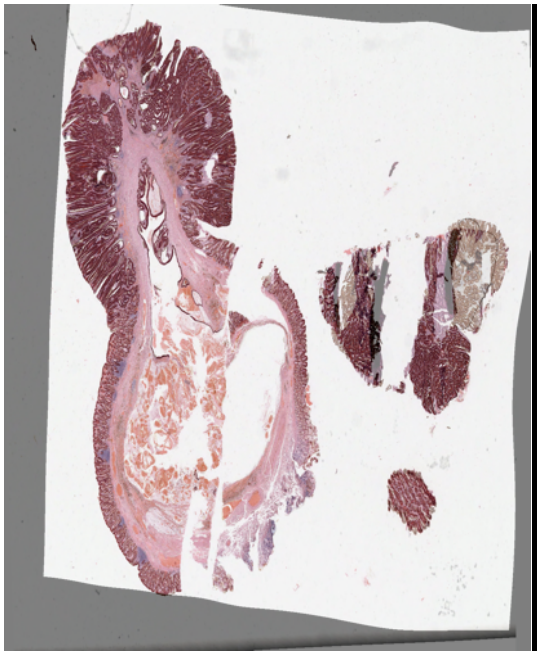
N



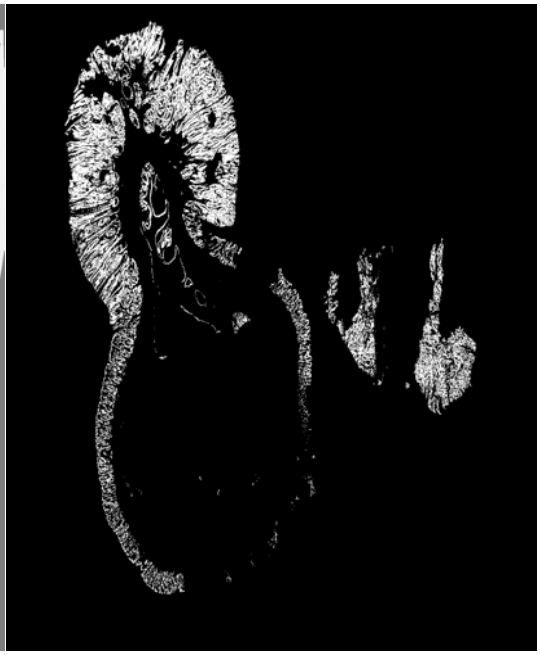
O



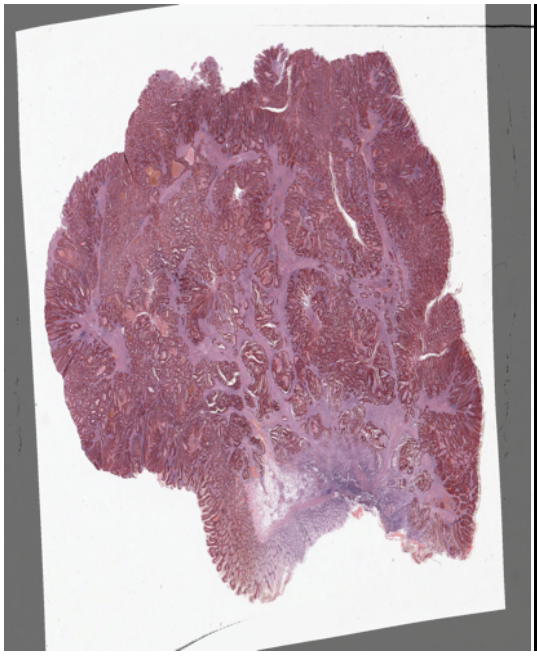
P



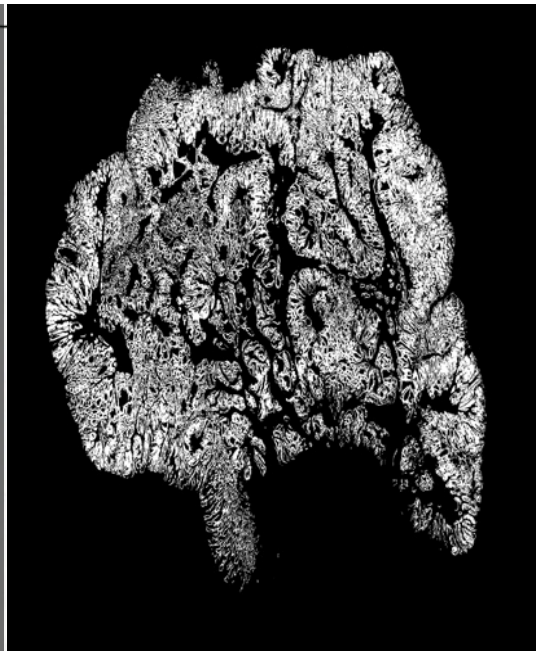
Q



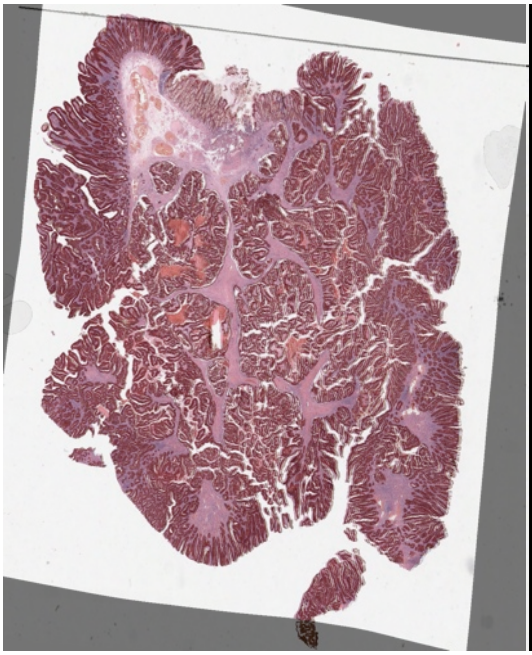
R



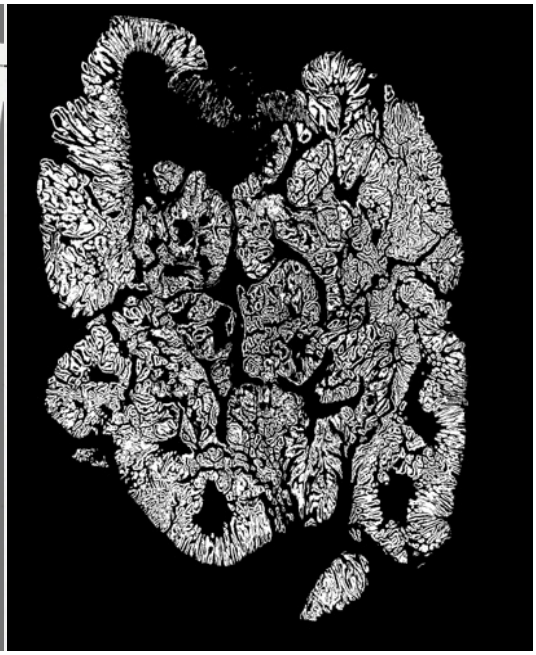
S



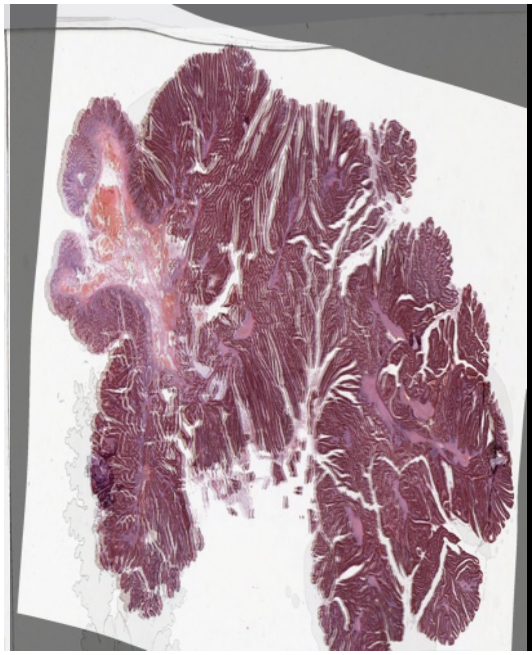
T



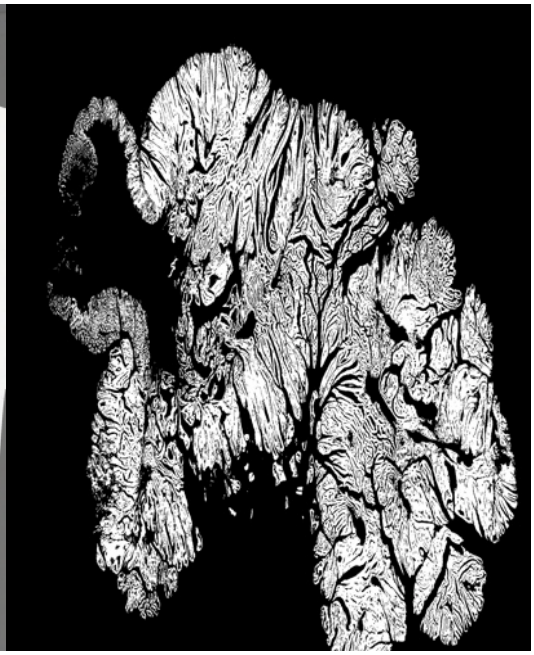
U



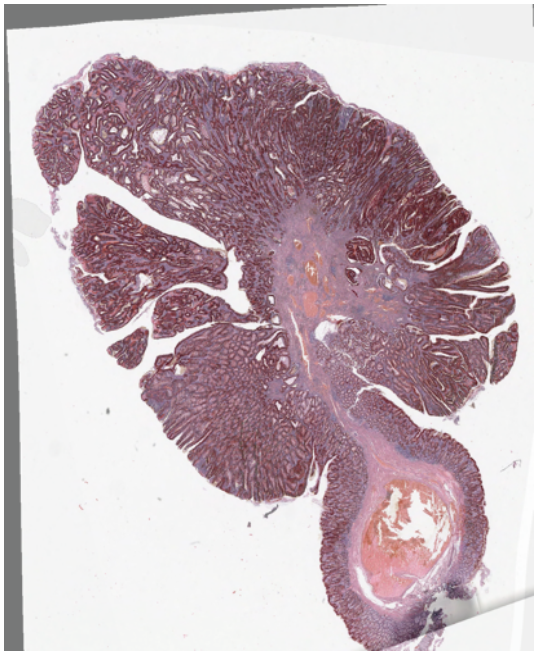
V



W

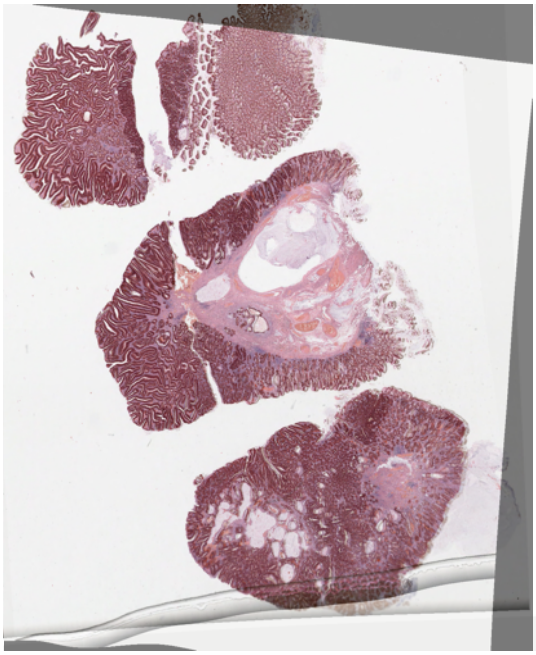


X

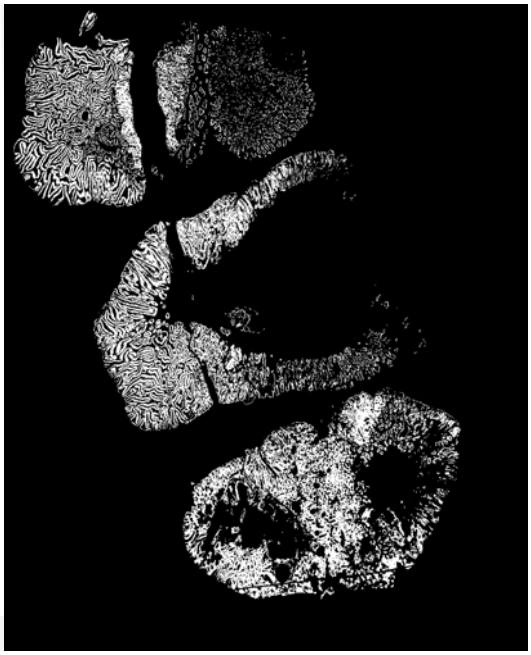


Y

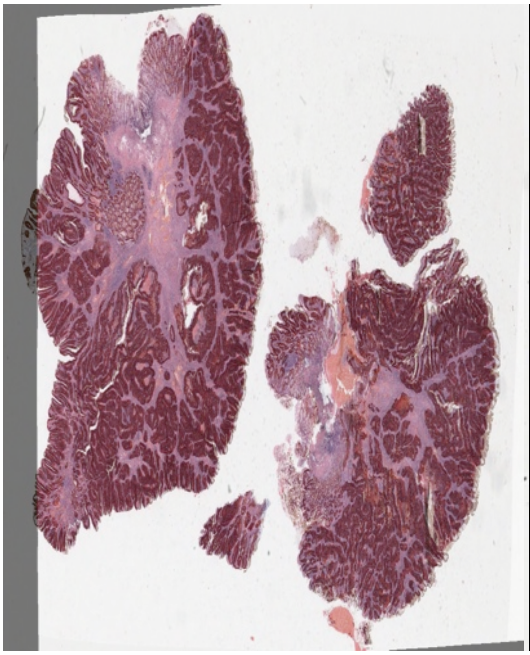
Z



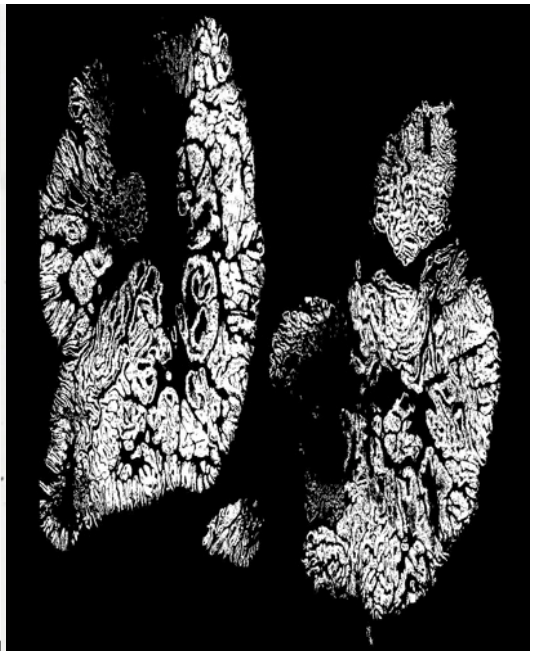
AA



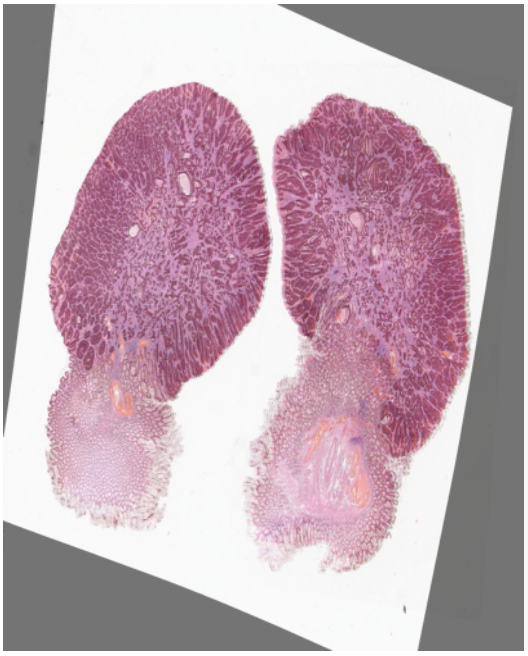
AB



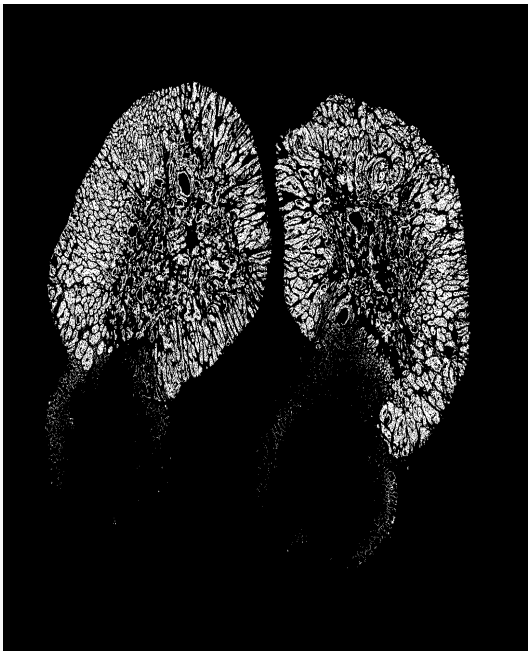
AC



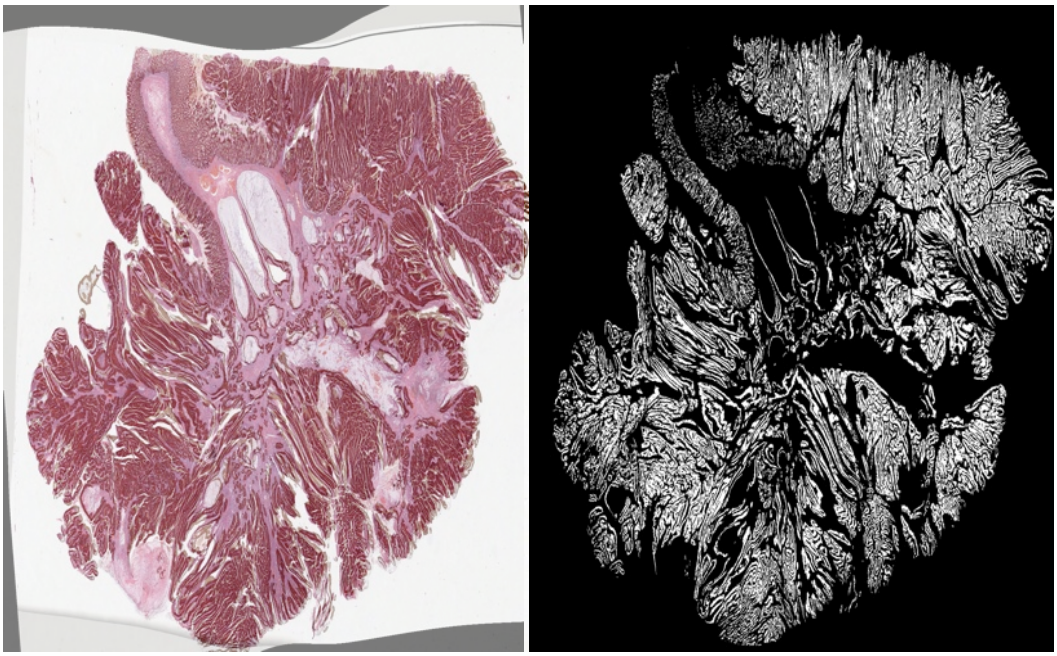
AD



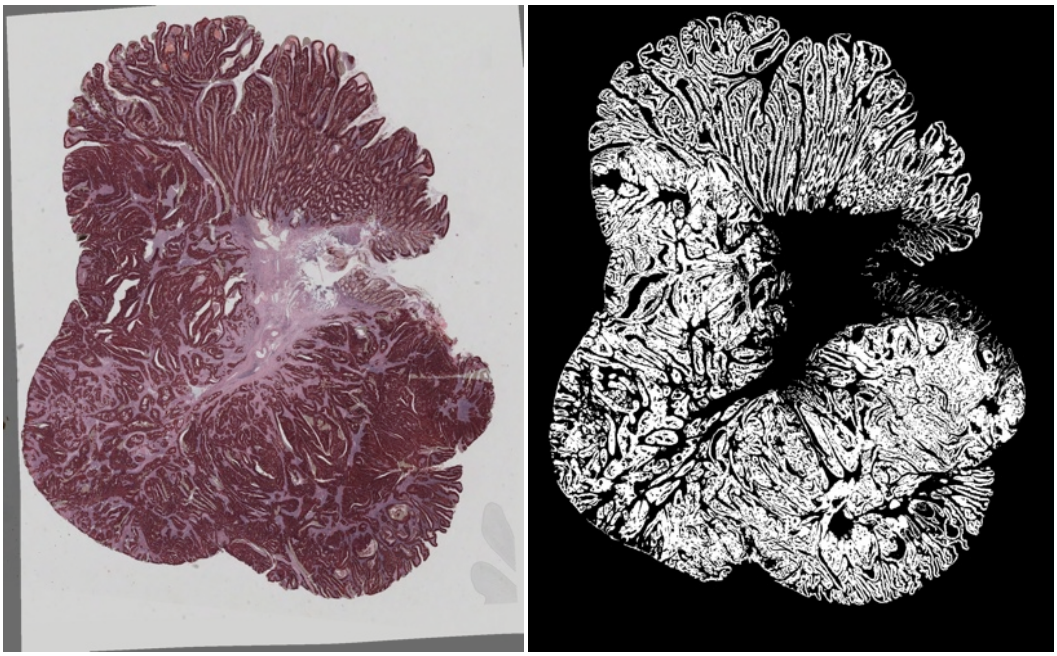
AE



AF

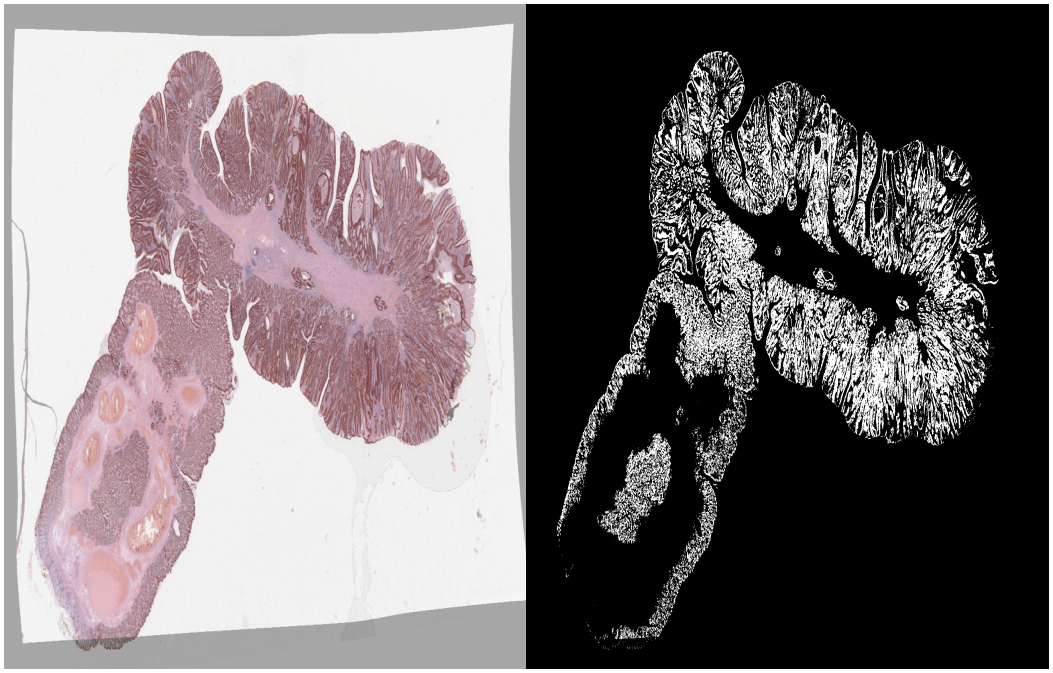


AG AH



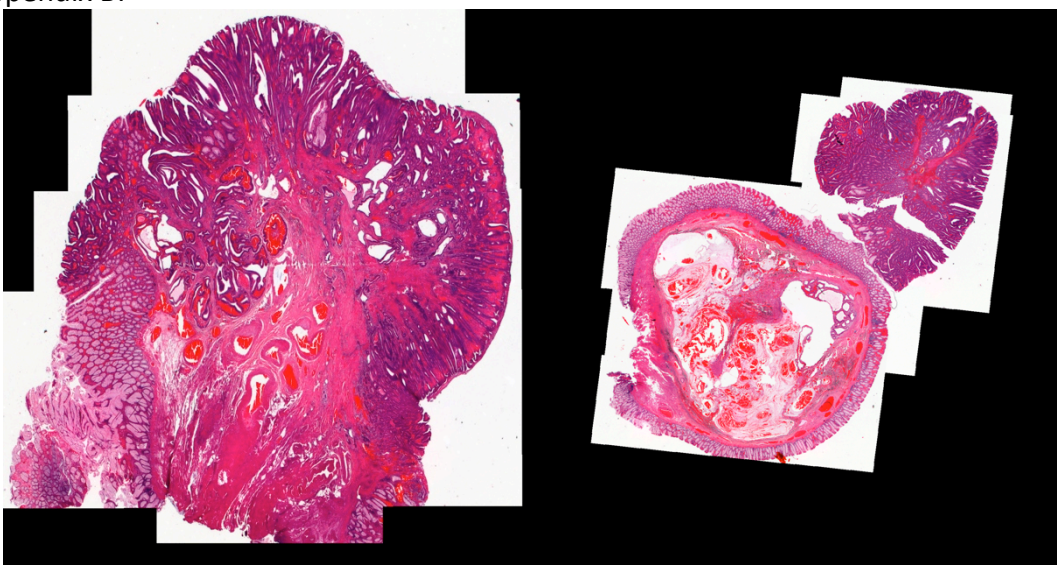
AI AJ

Appendix A. The first intermodally registered Haematoxylin and Eosin (H&E) and Cytokeratin (CK) images for all samples and their segmented counterparts.

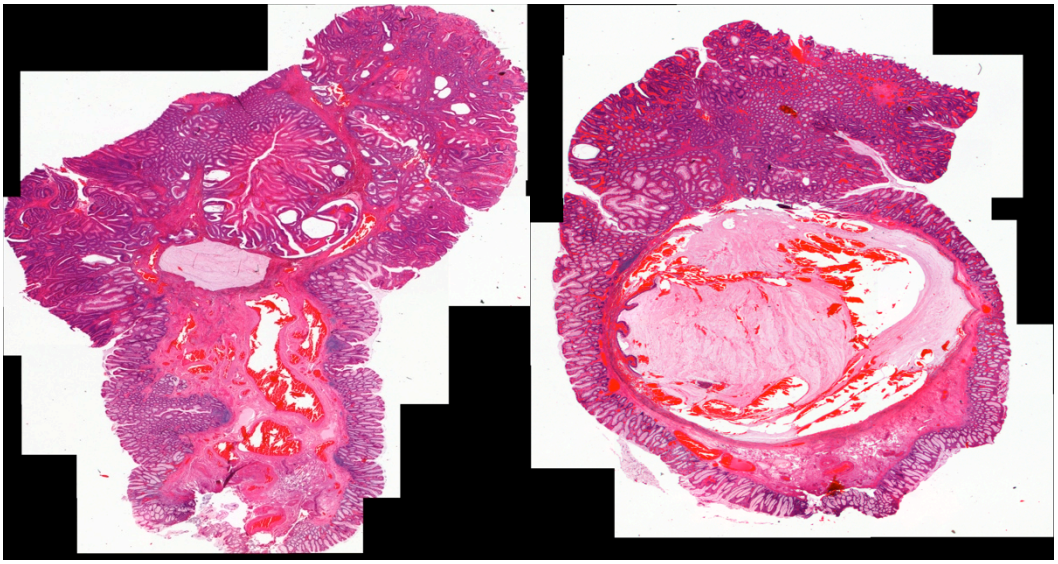


AK AL

Appendix B:



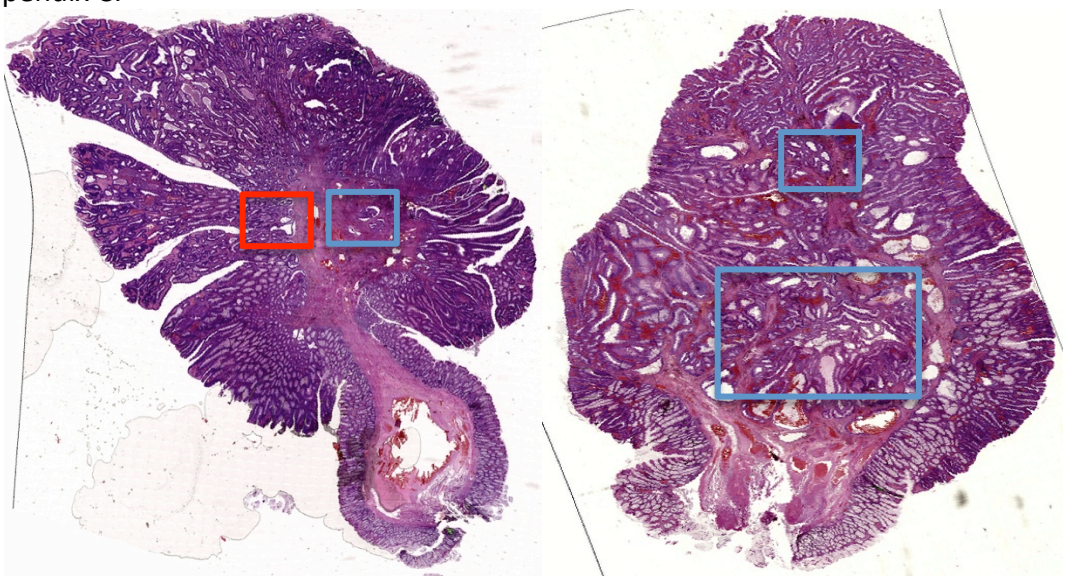
A B



C D

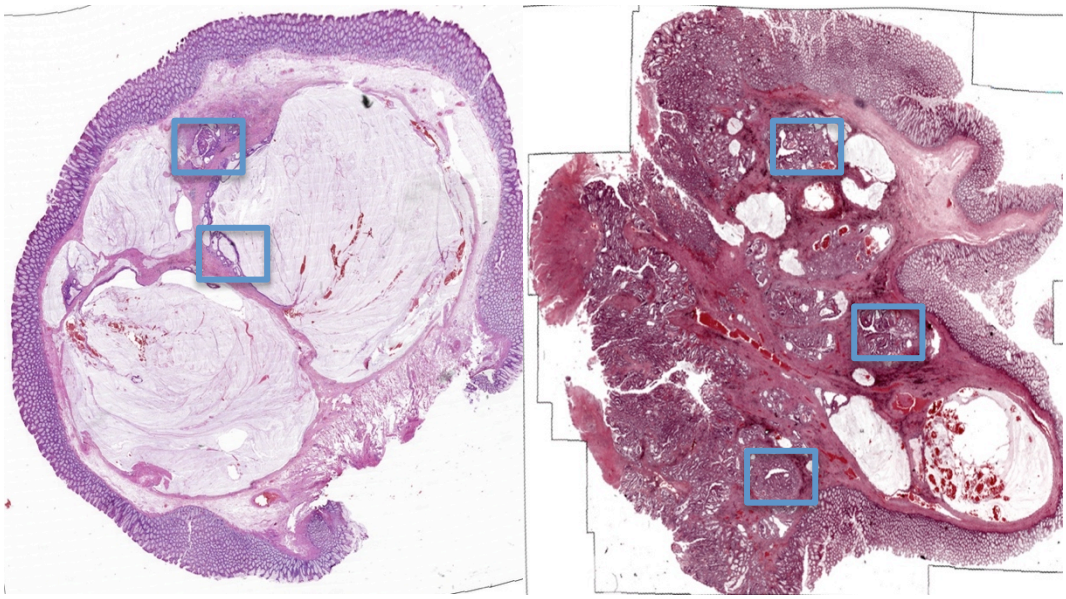
Appendix B: The H&E equivalents of the mucin segmentation images presented on the right of Figure 4.3.1.5.1. It can be noted in A that the mucinous regions for this sample are small in both number and size.

Appendix C:



A

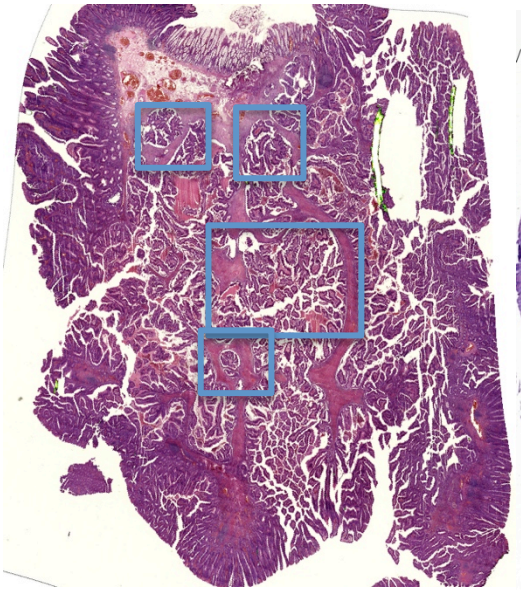
B



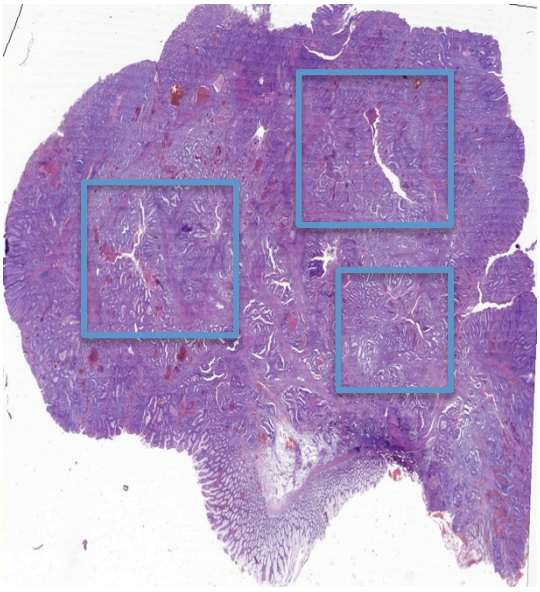
C

D

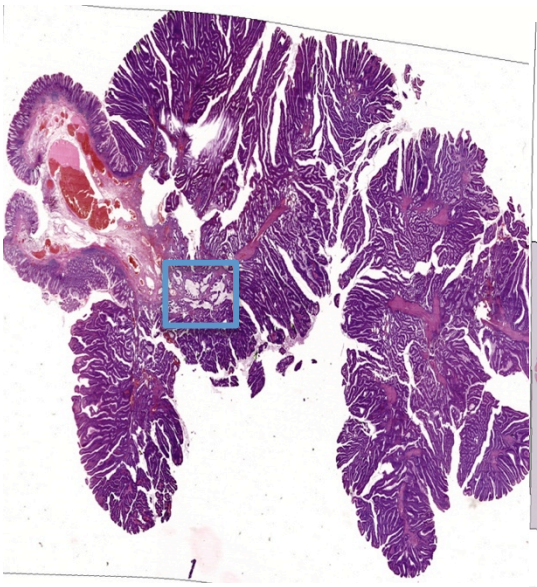
Appendix C: Part one. The red box represents regions that could not be sampled because of poor quality spectra.



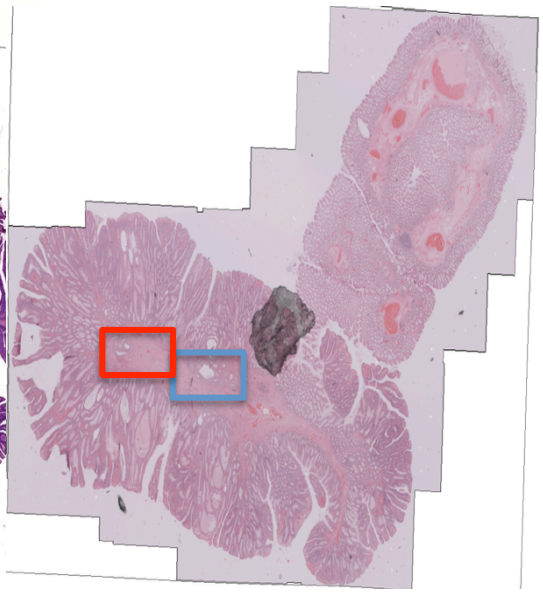
E



F

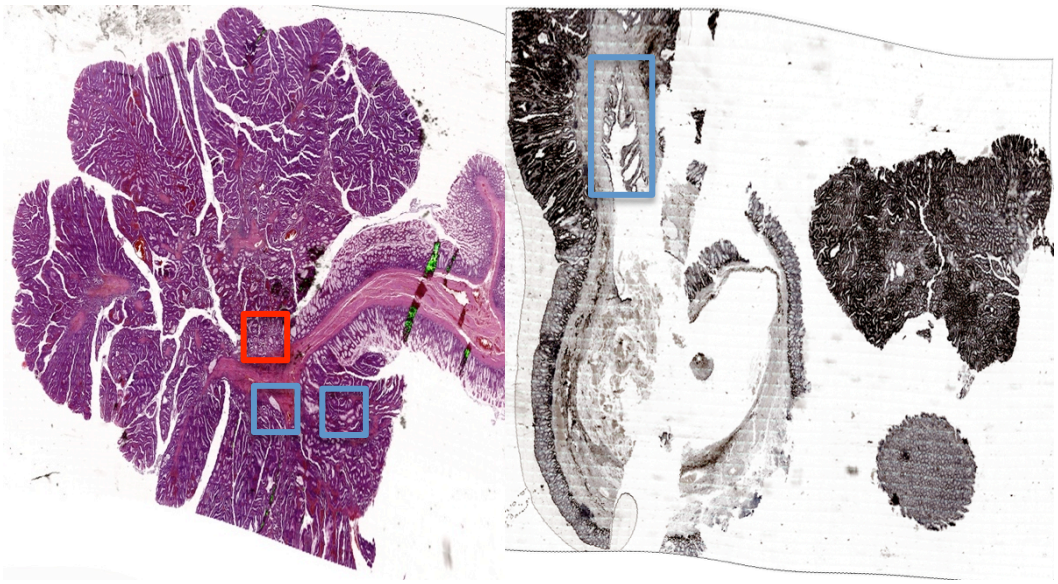


G

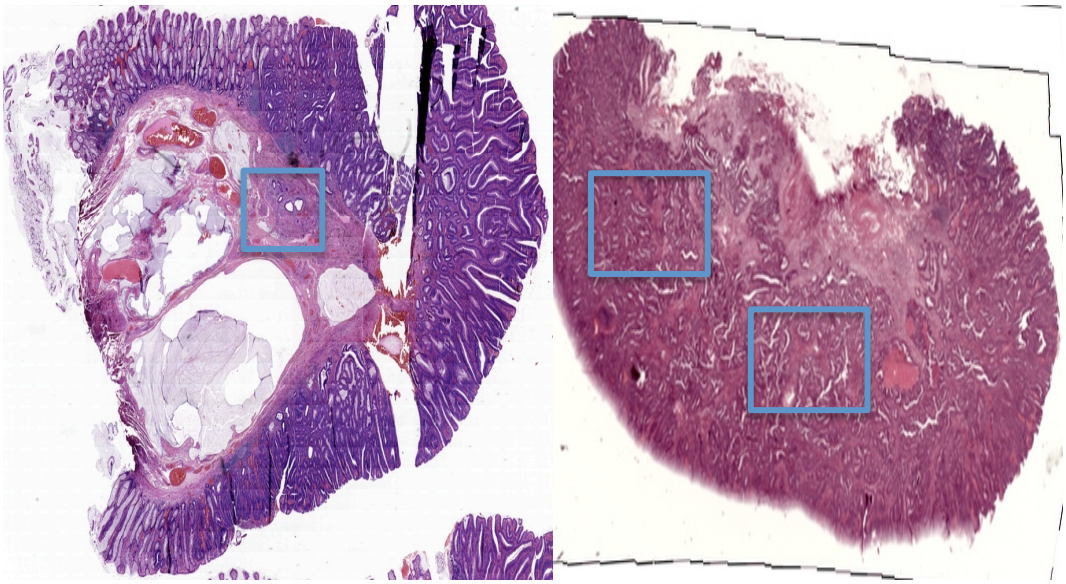


H

Appendix C: Part two. The red box indicates regions that could not be sampled because of poor quality spectra.

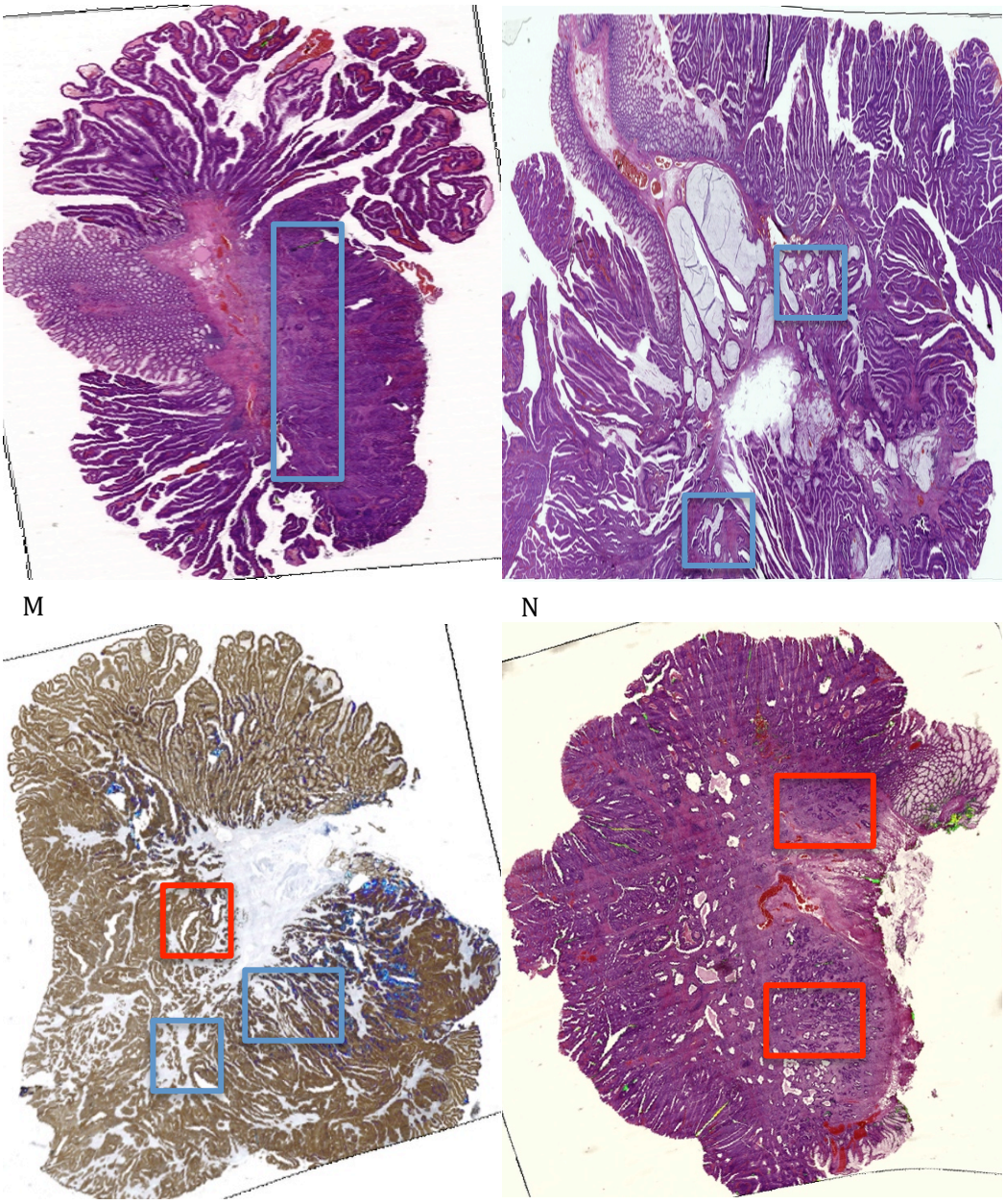


I J



K L

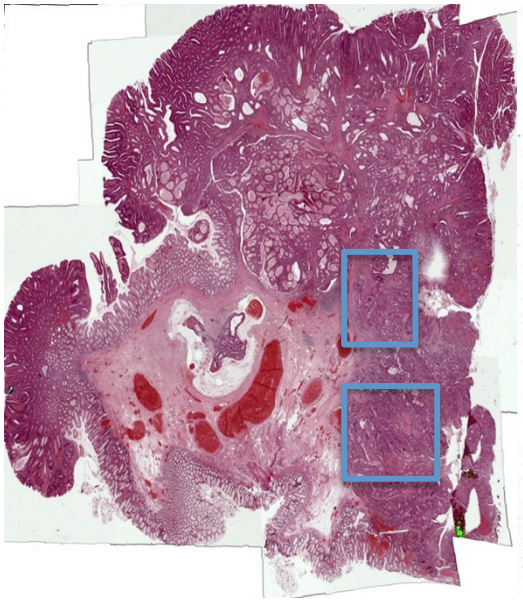
Appendix C: Part three. The red box represents regions that could not be sampled because of poor quality spectra.



M N

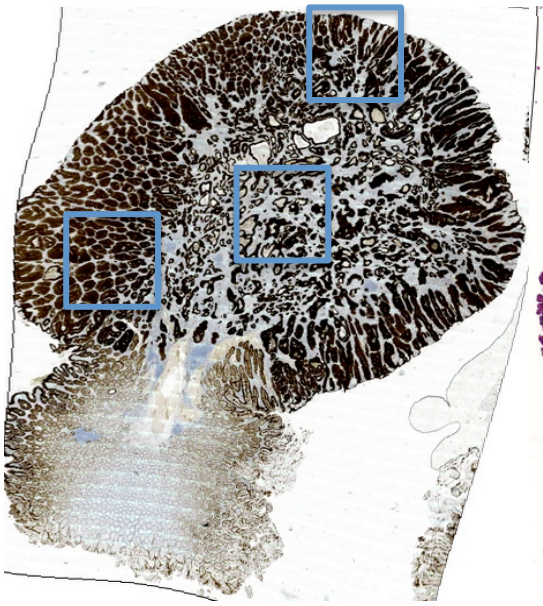
O P

Appendix C: Part four. The red boxes represents regions that could not be sampled because of issues relating to poor quality spectra.

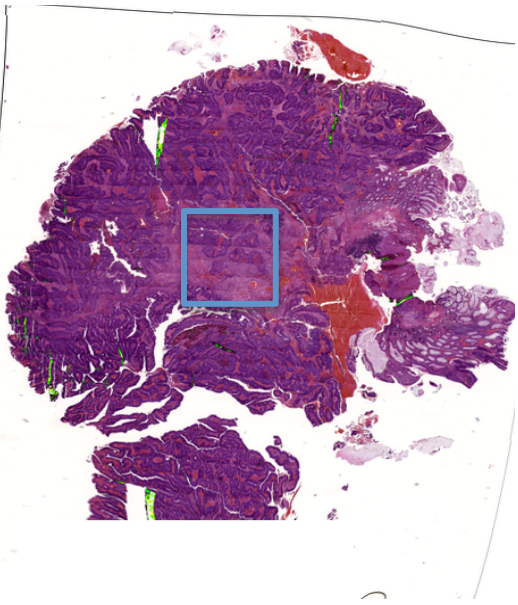


Q

R

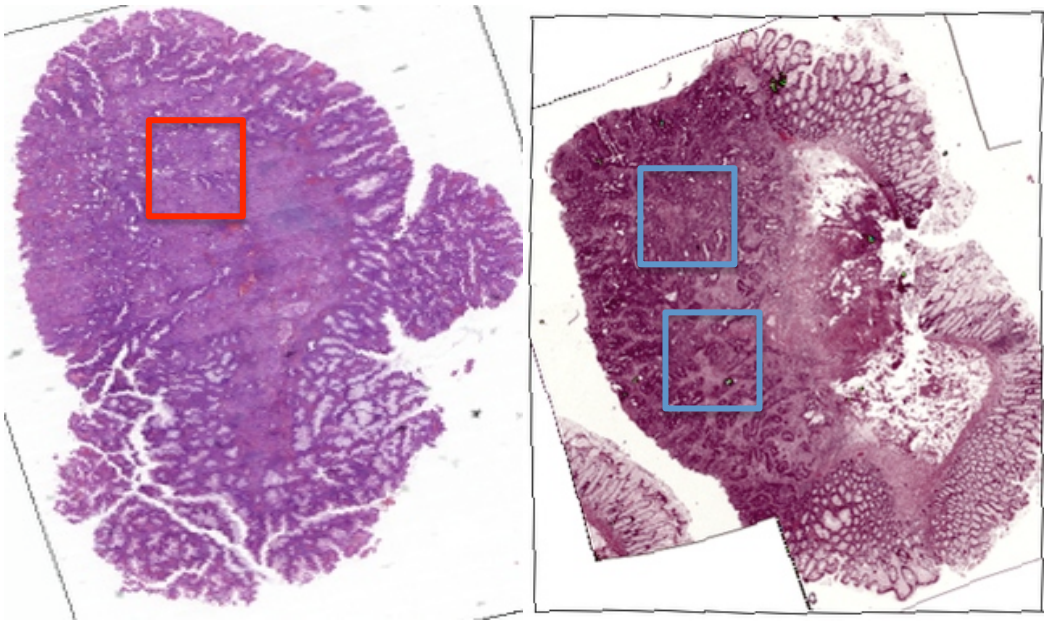


S



T

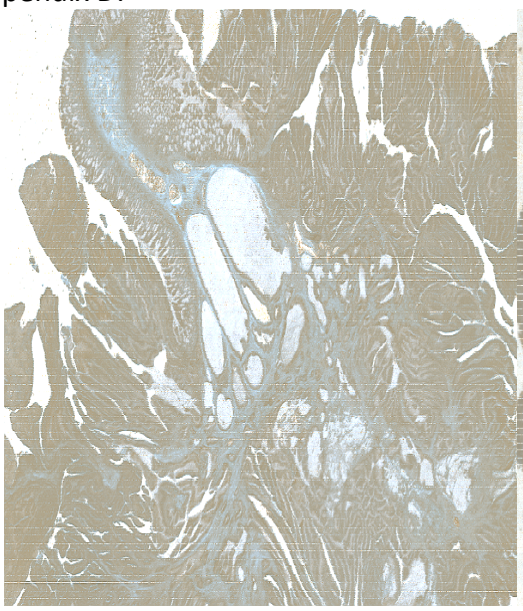
Appendix C: Part five.



U V

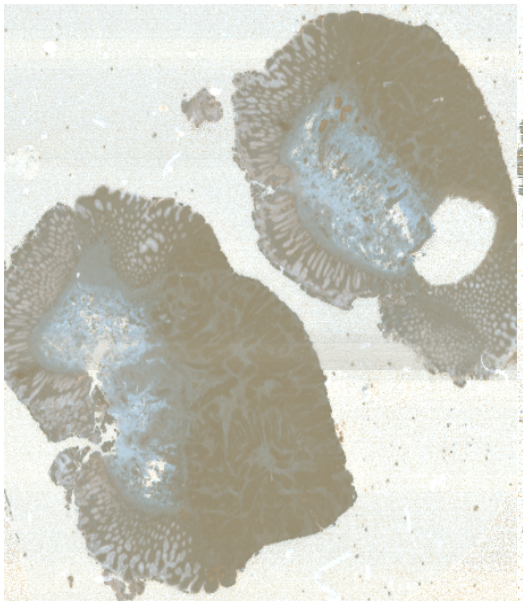
Appendix C: Part six. . The red boxes represents regions that could not be sampled because of issues relating to poor quality spectra.

Appendix D:



A (CK)

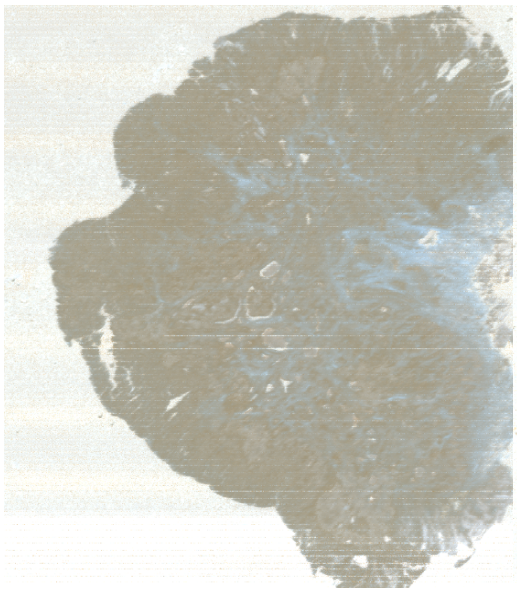
B (CK)



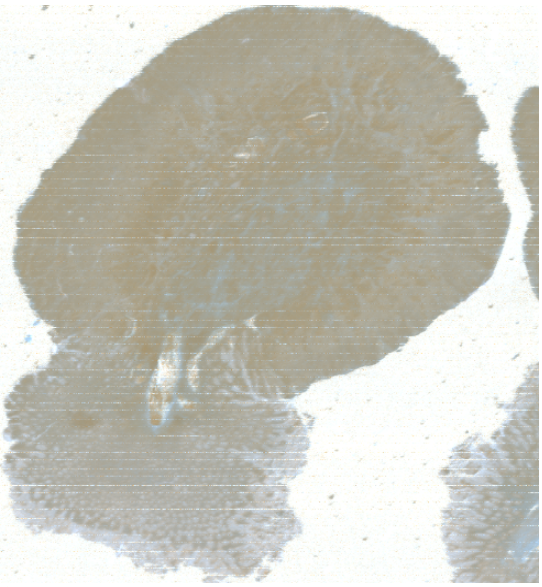
C (CK)

D (CK)

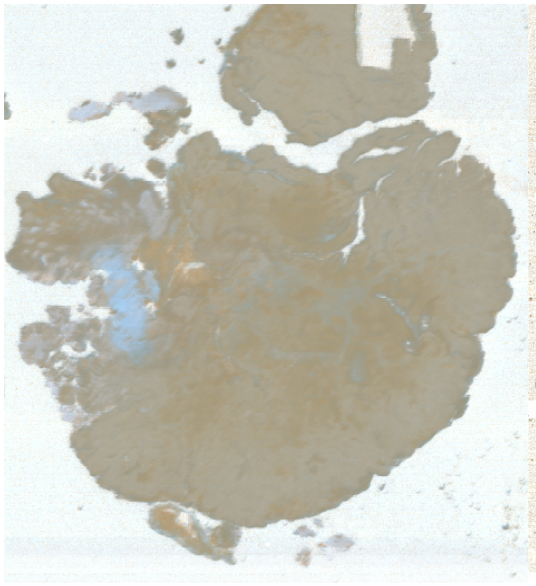
Appendix IV: Part one



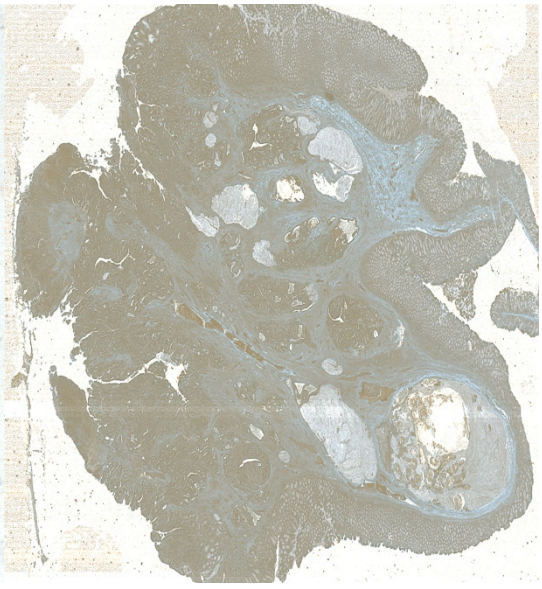
E (CK)



F (CK)

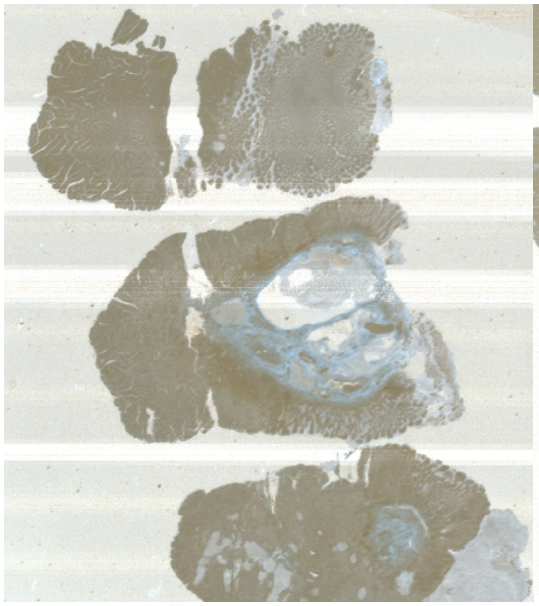


G (CK)



H (CK)

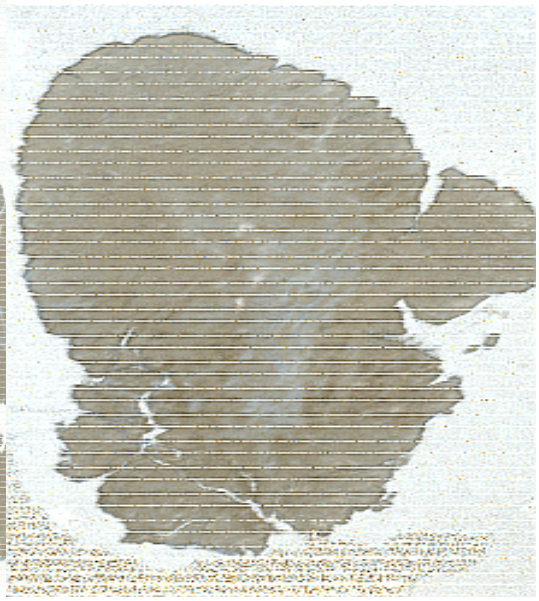
Appendix D: Part two. Note that no spectra were collected from image E(CK).



I (CK)

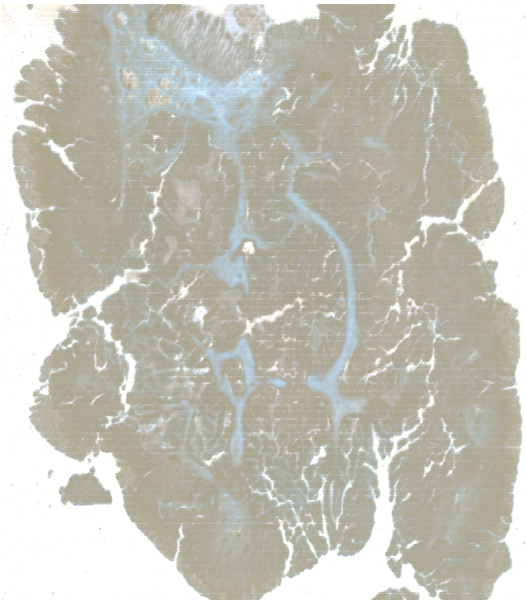
J (CK)

K (CK)

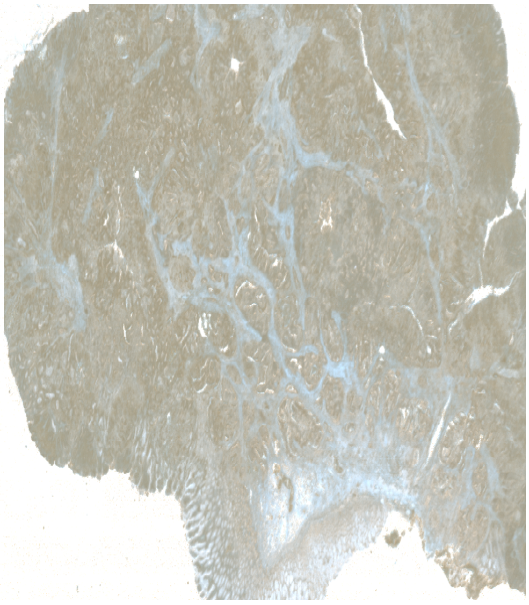


L (CK)

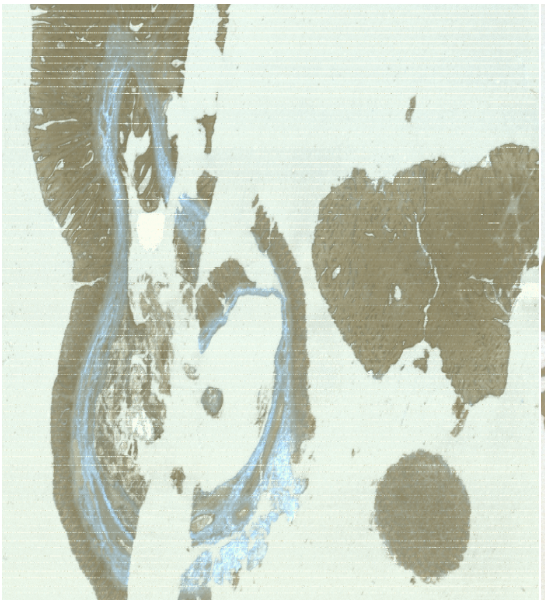
Appendix D: Part three. Please note that no spectra could be derived from L(CK)



M (CK)



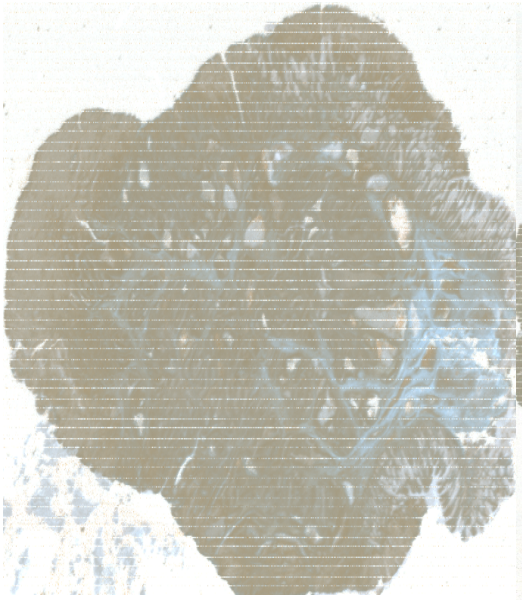
N (CK)



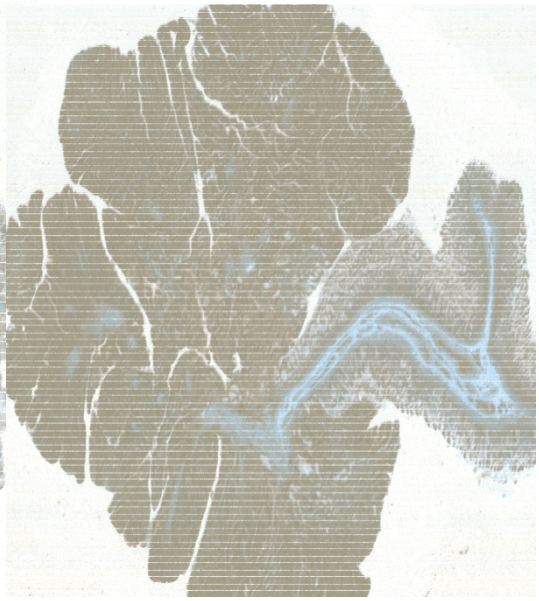
O (CK)

P (CK)

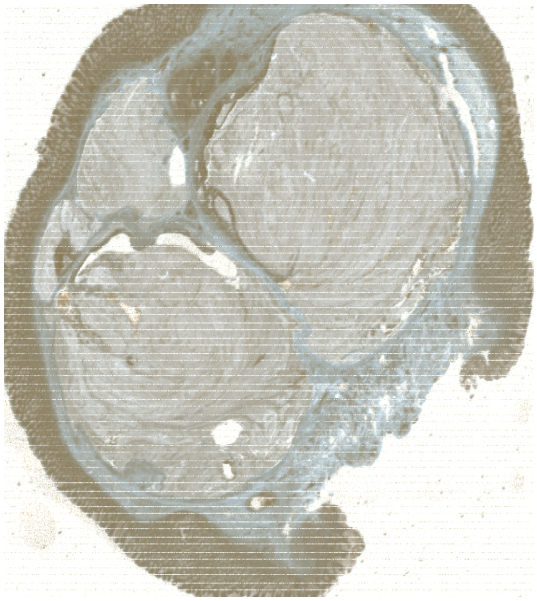
Appendix D: Part four



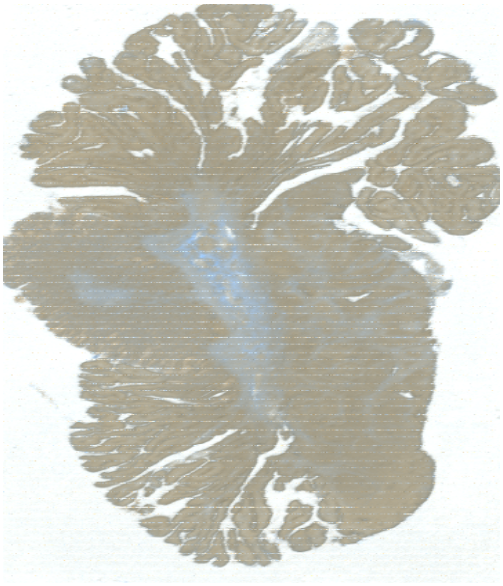
Q (CK)



R (CK)

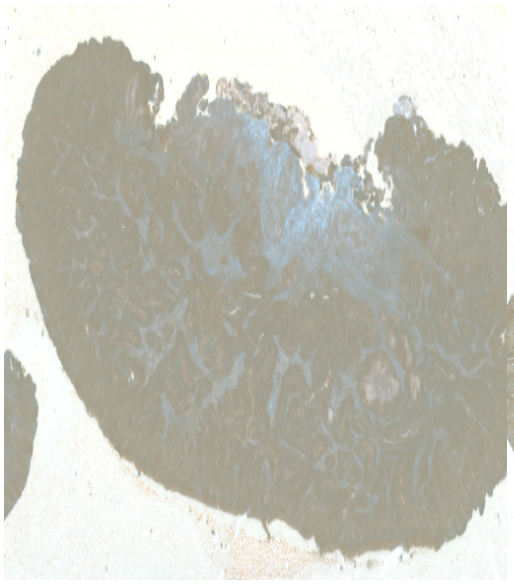


S (CK)

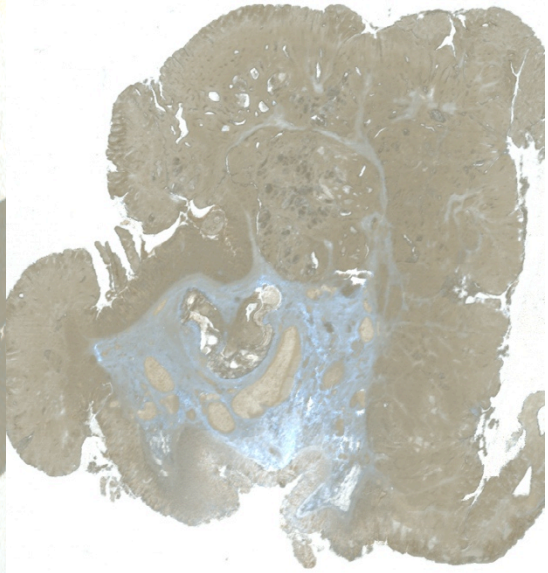


T (CK)

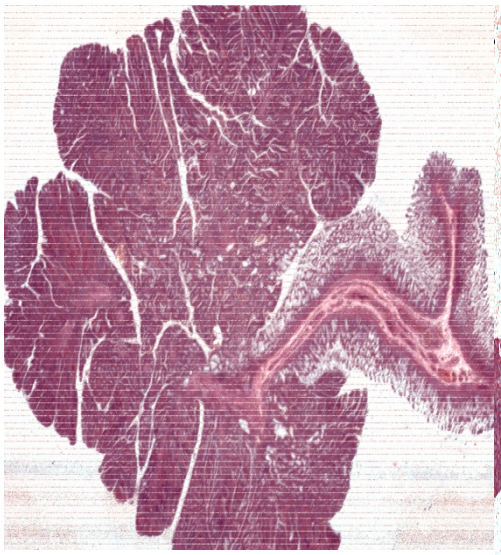
Appendix D: Part five.



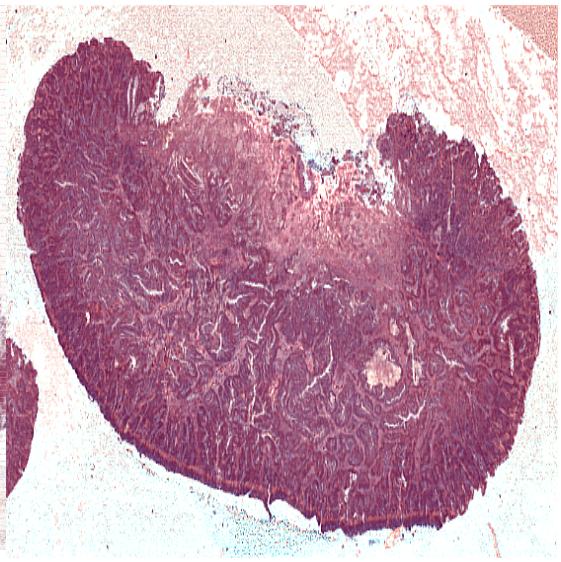
U (CK)



V (CK)

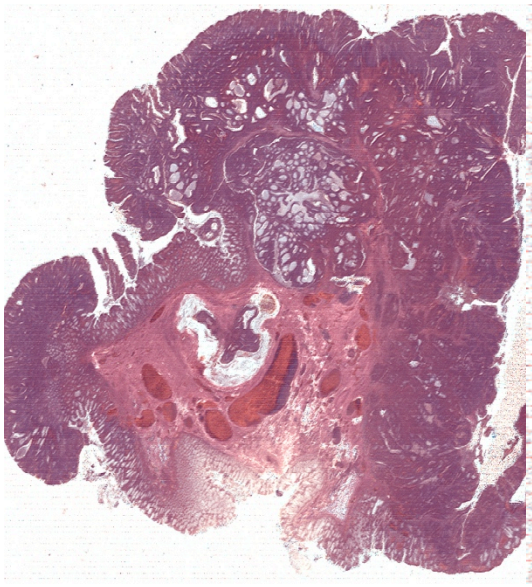


A (H&E)

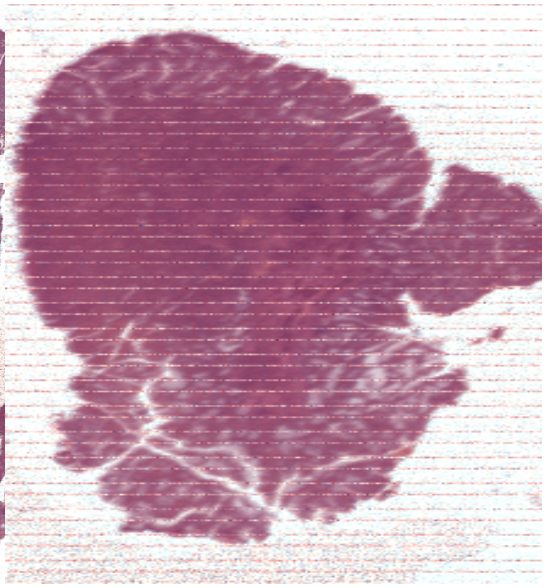


B (H&E)

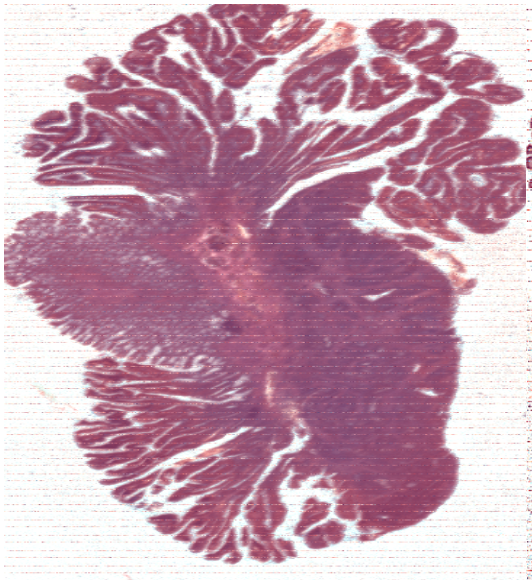
Appendix D: Part six**.**



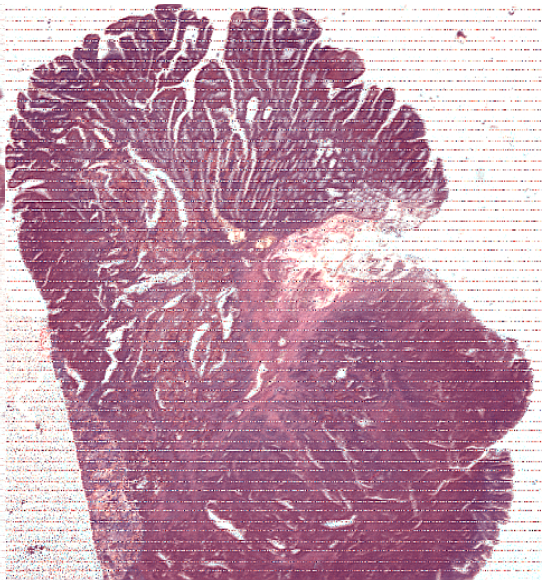
C (H&E)



D (H&E)



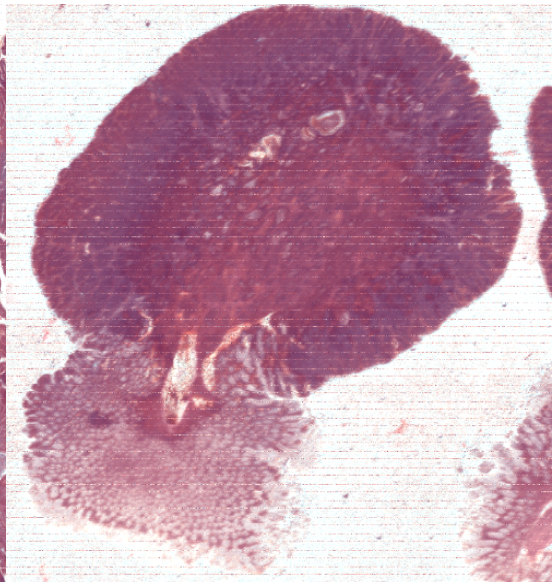
E (H&E)



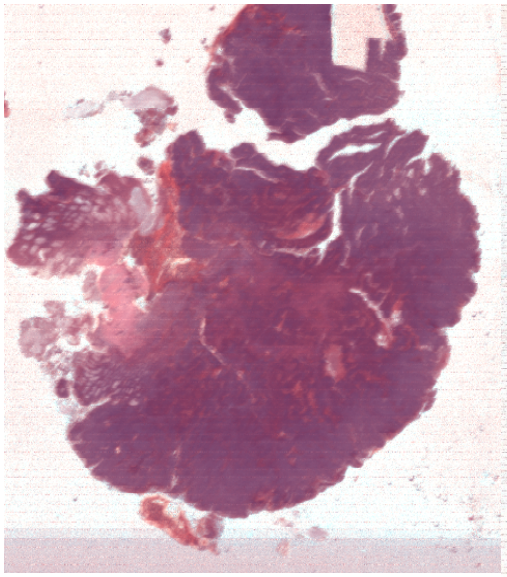
F (H&E)

Appendix D: Part seven. Note that no spectra were collected from image D(H&E)

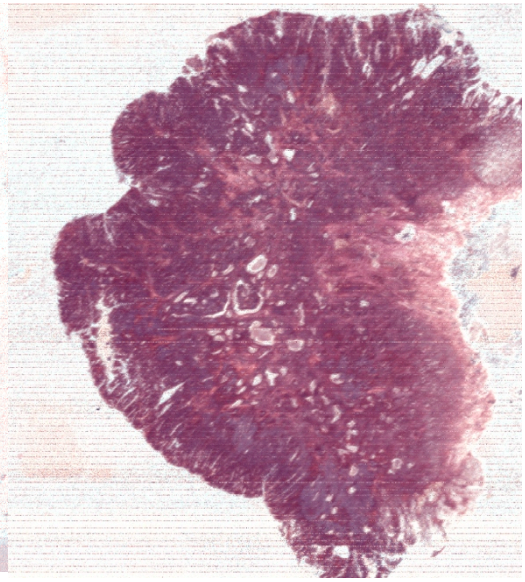
G (H&E)



H (H&E)



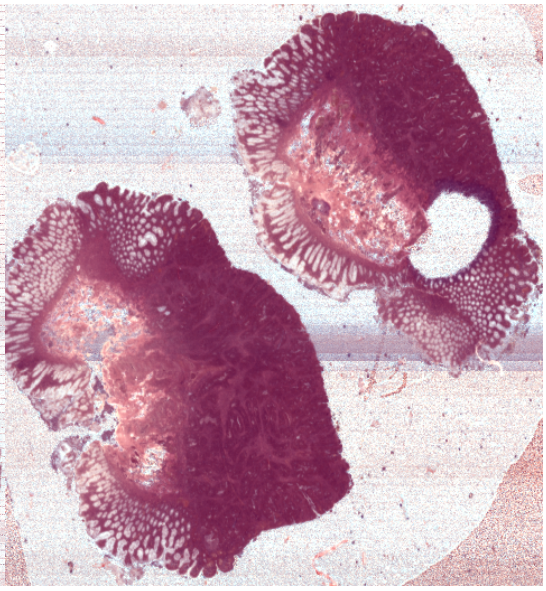
I (H&E)



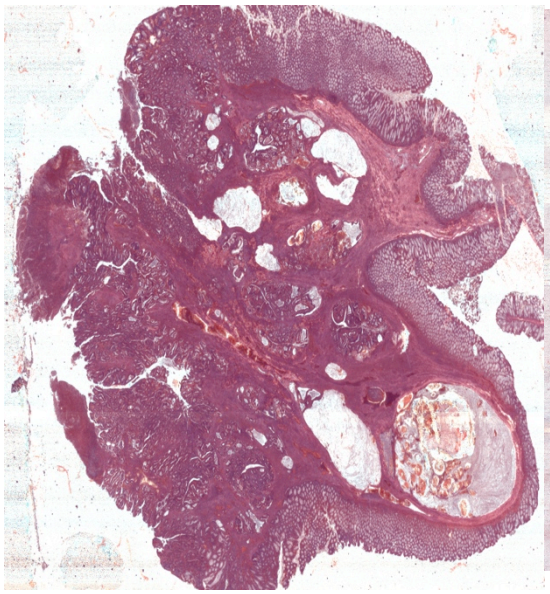
J (H&E)

Appendix D: Part eight. Please note that no spectra were derived from image J(H&E)

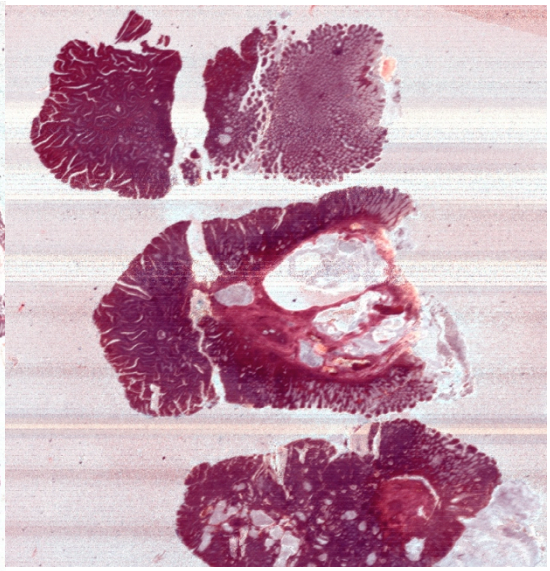
K (H&E)



L (H&E)

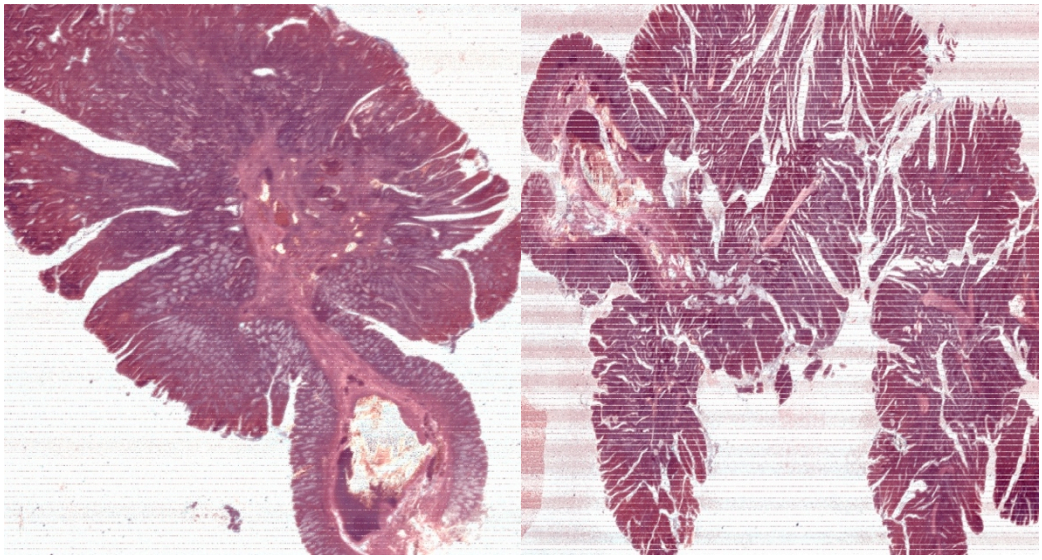


M (H&E)



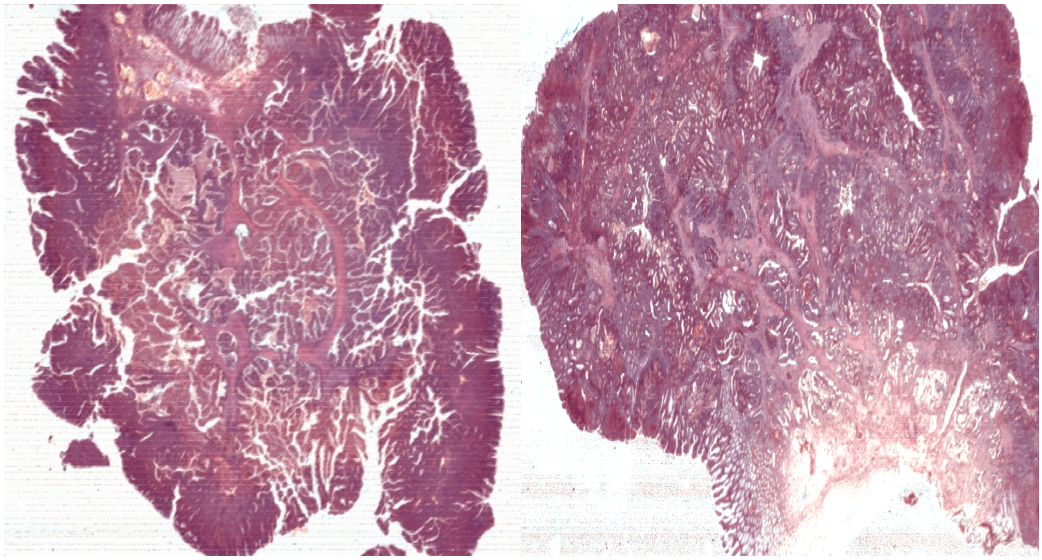
N (H&E)

Appendix D: Part nine.



N(H&E)

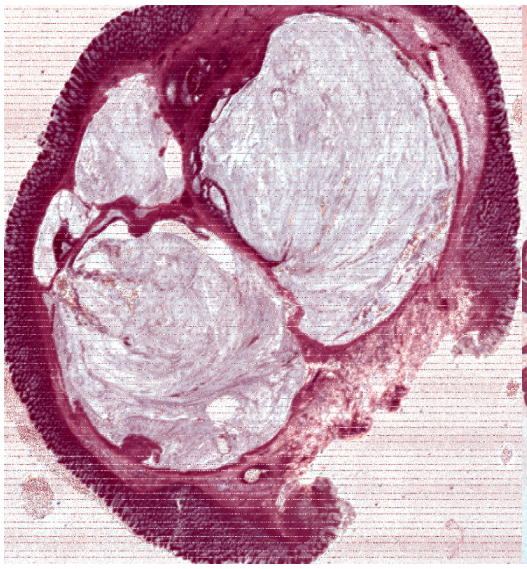
O (H&E)



P (H&E)

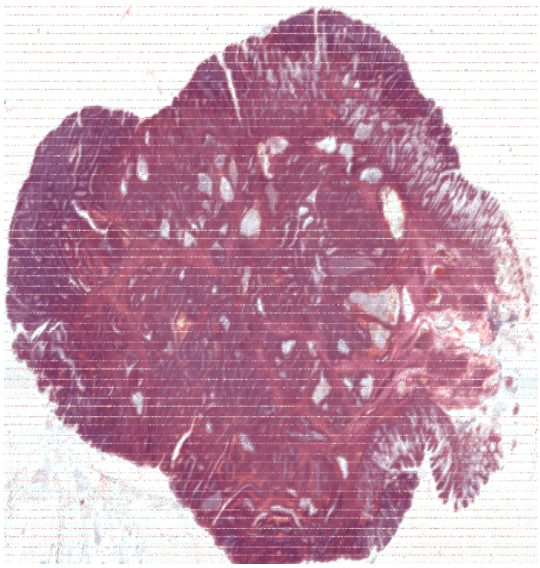
Q (H&E)

Appendix D: Part ten.



S(H&E)

T (H&E)

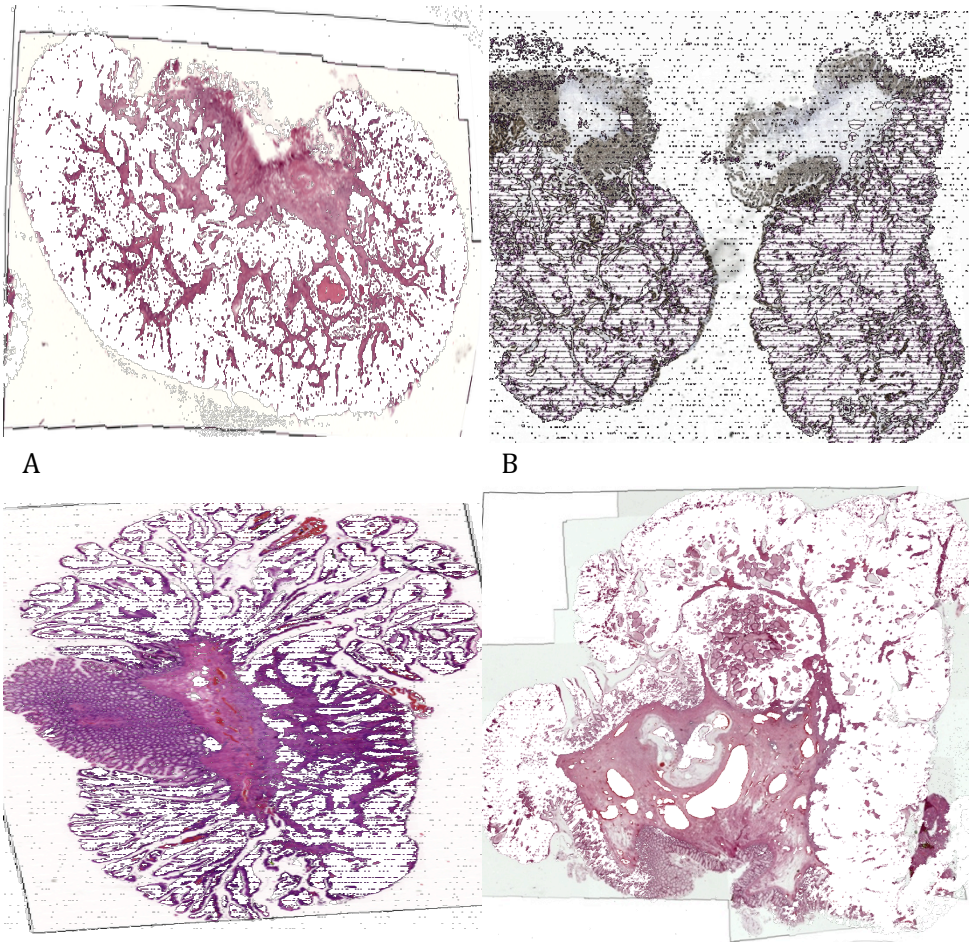


U (H&E)

V (H&E)

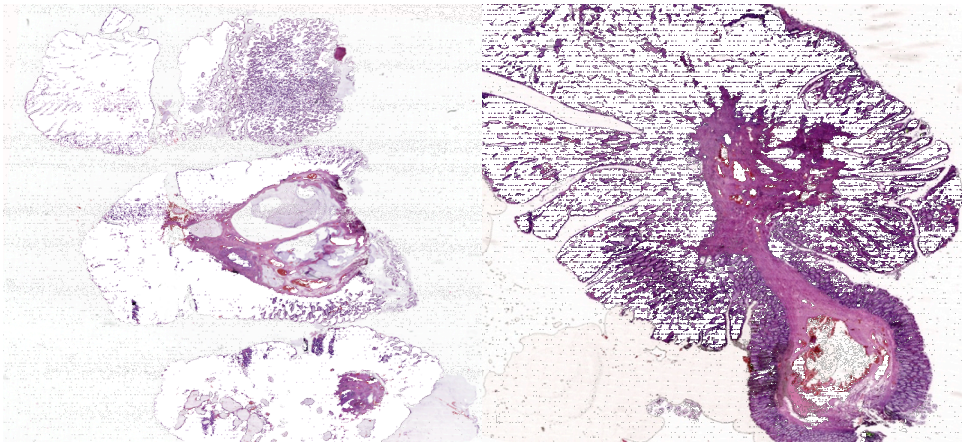
Appendix D: Part eleven.

Appendix E:

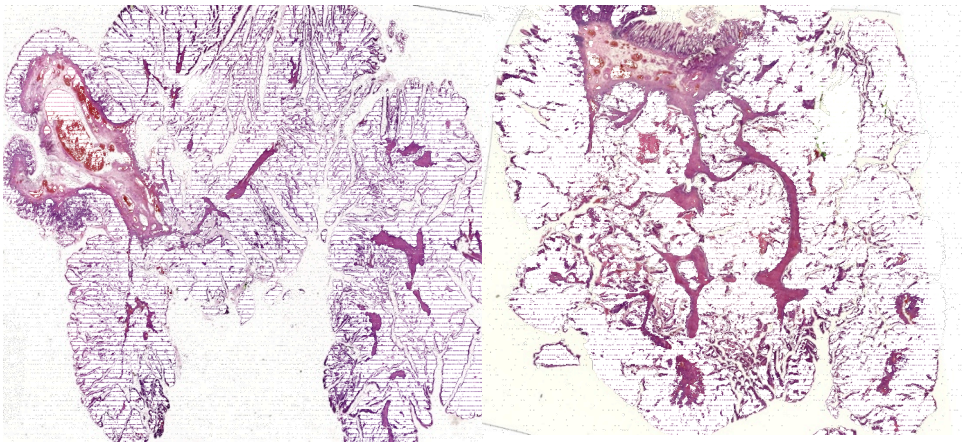


A B C D

Appendix E: Part one.

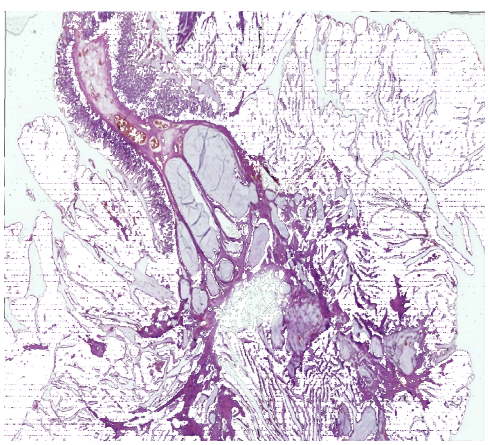


D E

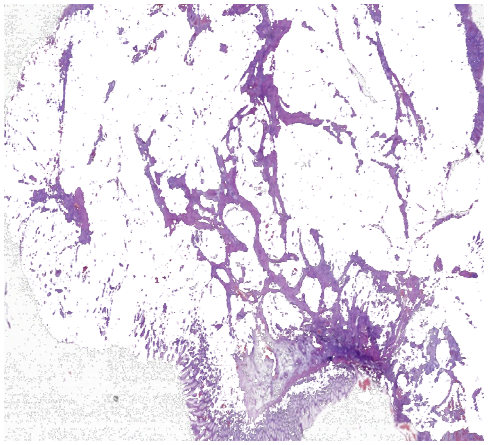


F G

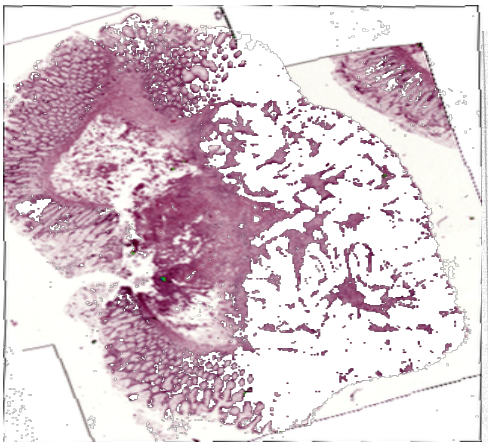
Appendix E: Part two.



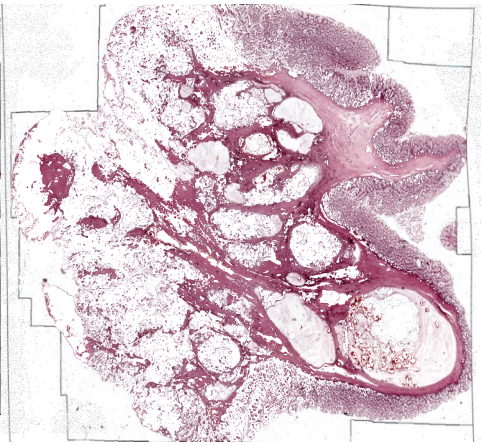
H



I

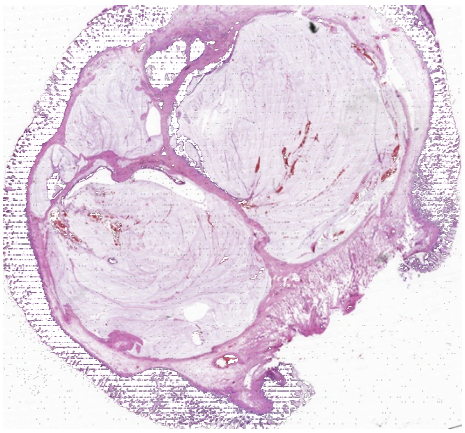


J

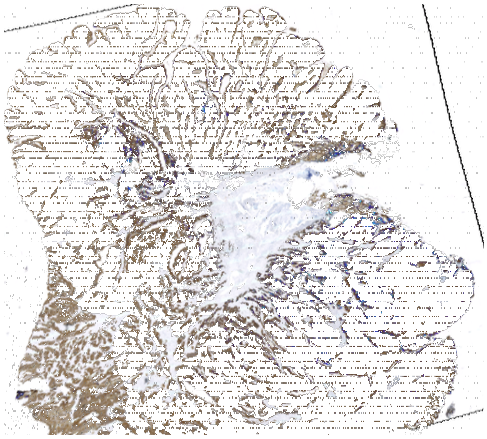


K

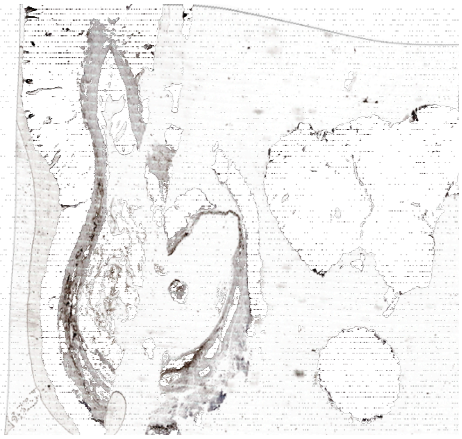
Appendix E: Part three.



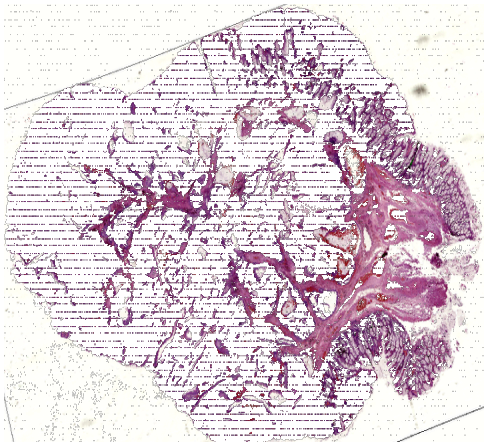
L



M

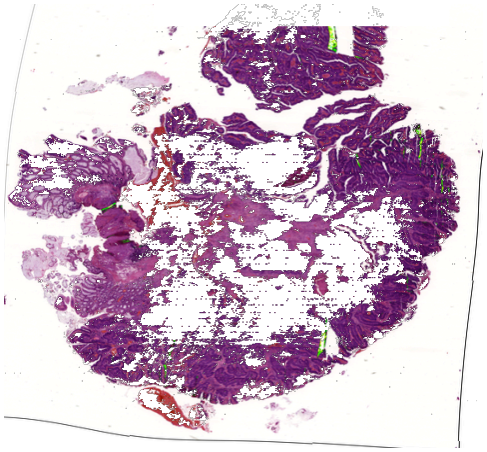


N

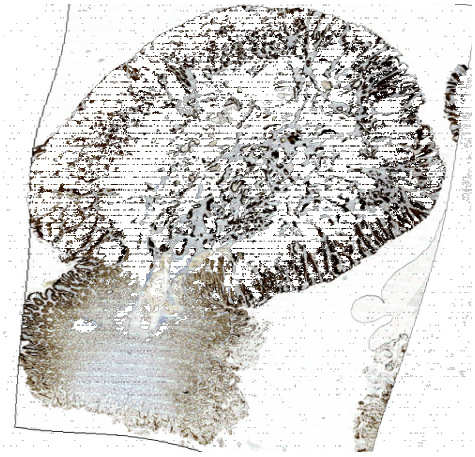


O

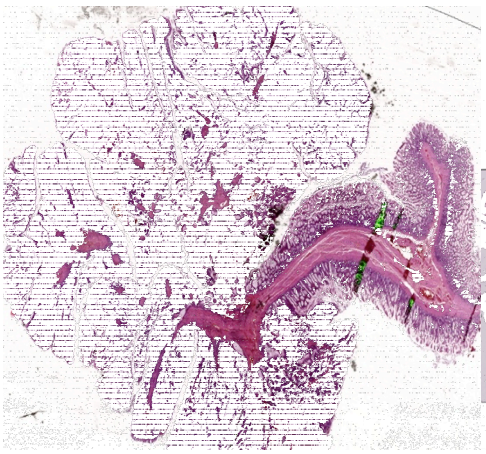
Appendix E: Part four.



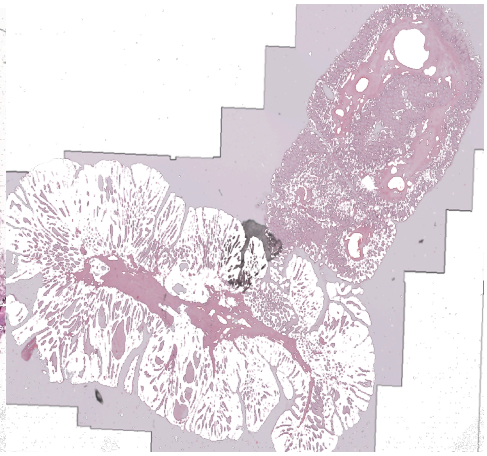
P



Q



R



S

Appendix E: Part five**.**

For the above images the black background has been removed and the segmented images have been overlain on their histological counterparts. From these images it can be noted that the registration between images was good and that the epithelium is generally being well segmented. However, the success of the segmentation is variable but epithelial regions can still be selected from all images**.**

Appendix F

| Movie Title | Pathology |
|---|---|
| **11_24307_2_seg** | EM |
| **06893seg** | Cancer |
| **075441seg** | Cancer that also contains EM |
| **093444seg** | Cancer |
| **101259seg** | EM |
| **113200seg** | EM |
| **113367SEG** | EM |
| **117117seg** | EM |
| **121566final** | EM |
| **123645seg** | Cancer |
| **1012414seg** | Cancer |
| **1021650seg** | EM |
| **1028668seg** | EM |
| **1034793seg** | Cancer |
| **1110694** | Cancer |
| **1119234SEG** | EM |
| **1124307seg** | EM |
| **1136972seg** | Cancer |
| **1231342seg** | EM |
| **additional** | EM |
| **additional2** | EM |
| **seg10183** | Cancer |

The pathological classification of the movies found within Electronic Appendix A.